KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Noise Robust Exemplar Matching for Speech Recognition and Enhancement

Emre Yılmaz

Supervisor:
Prof. dr. ir. Hugo Van hamme

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

May 2015

Noise Robust Exemplar Matching for Speech Recognition and Enhancement

Emre YILMAZ

Examination committee:
Prof. dr. ir. Paul Van Houtte, chair
Prof. dr. ir. Hugo Van hamme, supervisor
Prof. dr. ir. Dirk Van Compernolle, co-supervisor
Prof. dr. ir. Patrick Wambacq
Prof. dr. ir. Johan Suykens
Prof. dr. Bhiksha Raj
(Carnegie Mellon University, USA)
Prof. dr. Tuomas Virtanen
(Tampere University of Technology, Finland)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering

May 2015

Preface

Exploring the mysterious nature of speech under the roof of KU Leuven has been a great pleasure and the top challenge in the last years of my life. This challenge has been handled rather well mostly thanks to the huge love and support I received from my family, friends and colleagues. The inspiring and peaceful environment of Leuven, and Belgium in general, boosted our progress towards a novel noise robust automatic speech recognition approach. As of now, there exists a noise robust exemplar matching system for speech signals, the fundamentals of which are explained thoroughly in the following pages.

First and foremost, I would like to thank my supervisor, Prof. Hugo Van hamme, for his constant support and guidance. I have learned immensely from his way of approaching the problems we have encountered and this work could have never been done without his brilliance. I will always remember and appreciate our long discussions during which I had time to establish my own research perspective.

Many thanks go to my co-supervisor, Prof. Dirk Van Compernolle, and collaborators, Dr. Jort Gemmeke and Deepak Baby, for their contributions and valuable comments. I would like to express my gratitude towards Prof. Patrick Wambacq, Prof. Johan Suykens, Prof. Tuomas Virtanen and Prof. Bhiksha Raj for being a member of my jury. Their remarks made this thesis more readable and informative. Many hugs to past and present members of the Speech group, they were always very friendly and helpful towards my technical and non-technical problems.

Leuven is also a very lively and social city offering marvelous opportunities in various aspects of life. Going out and meeting up with friends on Oude Markt was a great way to relax and get prepared for the next day. I would like to thank my friends from the PhD Society Leuven for our cheerful meetings to organize activities to support the young academic staff in KU Leuven. Thinking of our badminton games with Xueru and Joris will always put a smile on my

face. Cheers to all friends who were with me during the pintje moments.

I should also mention my old friends from Turkey who have been living far away but still always standing next to me. Elif, many thanks for listening to me patiently at all times. Great to know that time difference cannot stop our contact. Greetings to my friends living in other parts of the world, Anıl, Can, Çağrı, Deniz, Emrah, Erhan, Gökhan, Onur, Sezer and Zeynep who stopped by at every possible occasion to share the joy.

Special thanks to my brother, Erdem, who moved to Leuven for his philosophy studies in 2011 and always was the closest company from the very beginning. I will never forget our long walks in the streets of Leuven and film-making attempts with the help of our friends Joris and Giovanni.

I do not know how to express my feelings with words, but during my PhD years in Leuven, I have also met my other half, Jeanette. Without her love, support and deep sympathy, I would have lost a significant portion of my energy, happiness and strength. Thanks for being my partner and sharing the life with me.

Son olarak beni her zaman koşulsuz olarak destekleyen anneme ve aileme şükranlarımı sunmak isterim. Bu tezin tamamlanmasındaki katkılarınızdan dolayı sizlere çok teşekkür ederim. Oma, Trudy en Jan, jullie zijn de liefste en warmste schoonfamilie ooit. Hartelijk dank voor de voortdurende steun tijdens mijn doctoraat.

To my dear father, Mehmet Ali Yılmaz...

Abstract

Automatic speech recognition (ASR) systems aim to map speech signals onto phonetic content, text, semantics and paralinguistics. This is achieved by learning some reference models for each acoustic unit in the speech feature space and comparing these models with the unknown test segments. Most ASR systems summarize the training data in an – often statistical – acoustic model such that multiple hypotheses can be evaluated on the test data and be compared such that the best hypothesis can be selected in a search procedure. By contrast, in template or exemplar based approaches, the test data is compared against *labeled* speech segments, i.e. the model is the training data.

The performance of the ASR systems is hindered by the background noise, i.e. the undesired signals that are also captured during the recording of the test speech signal. Therefore, a significant amount of research effort has been devoted to increase the noise robustness of the conventional ASR systems. The techniques proposed for improved noise robustness devise similar strategies such as using noise-immune speech features, enhancing the noisy signal or features prior to the recognition phase or noise compensated models to reduce the mismatch between the training and testing conditions. A class of noise robustness techniques that is particularly relevant to this thesis uses *unlabeled* speech and noise exemplars to linearly decompose noisy speech segments of fixed duration into a speech and noise component in order to enhance their representations.

In this work, we combine both template approaches described above: both the recognition and the noise reduction are based on the same exemplars. This results in a novel noise robust ASR scheme that uses speech and noise exemplars to model noisy speech. The exemplars are associated with a single speech unit similar to the traditional template matching ASR systems. The fundamental noise modeling problem of the template matching has been remedied by adopting a sparse representation formulation in which the exemplars are linearly combined to approximate noisy speech segments.

The proposed framework has been applied to the small vocabulary task of the 2nd CHiME Challenge and to the AURORA-2 database. The results on these small vocabulary ASR tasks have demonstrated the feasibility of the proposed technique. Moreover, the proposed scheme has been enriched in other aspects such as an accurate noise dictionary design procedure, data selection experiments for reduced computational complexity, embedding time warping to match feature sequences of different durations and a more flexible divergence metric for improved noise robustness. Finally, a single-channel speech enhancement system is proposed based on the same model and the recognition performance is evaluated using the enhancement system in the front-end of a conventional ASR system with Gaussian mixture models.

Abbreviations

AB	Alpha-Beta
AFE	Advanced Front-End
ANES	Active Noise Exemplar Selection
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
CD	Context Dependent
CDHMM	Continuous Density Hidden Markov Model
CMN	Cepstral Mean Normalization
CR	Collinearity Reduction
DCT	Discrete Cosine Transform
DDHMM	Discrete Density Hidden Markov Model
DNN	Deep Neural Network
DTW	Dynamic Time Warping
EM	Expectation Maximization
FE	Feature Enhancement
FST	Finite State Transducer
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ID	Interpolative Decomposition
KLD	Kullback-Leibler Divergence

LASSO	Least Absolute Shrinkage and Selection Operator
LVCSR	Large Vocabulary Continuous Speech Recognition
MAP	Maximum A Posteriori
MFCC	Mel-Frequency Cepstral Coefficients
MIDA	Mutual Information Discriminant Analysis
MS	Modulation Spectrogram
N-REM	Noise Robust Exemplar Matching
NN	Neural Network
NSC	Non-negative Sparse Coding
OM-LSA	Optimally-Modified Log-Spectral Amplitude estimator
PESQ	Perceptual Evaluation of Speech Quality
PLP	Perceptual Linear Prediction
RA	Recognition Accuracy
RASTA	Relative Spectra
SC	Sparse Classification
SDR	Signal-to-Distortion Ratio
SNR	Signal-to-Noise Ratio
SR	Sparse Representations
STFT	Short-Time Fourier Transform
SVD	Singular Value Decomposition
VAD	Voice Activity Detection
WER	Word Error Rate

List of Symbols

$\mathbf{O}$	The observed speech signal
o_t	The observed frame with time index t
$\mathbf{W}$	A word sequence
w_l	A word in the word sequence with index l
$\mathbf{Q}$	An HMM state sequence
q_t	An HMM state with time index t
Φ_i	Initial state probability of state i
$\mathbf{A}_{ij}$	Transition probability from state i to state j
$b_i(\mathbf{o})$	The emission probability of frame $\mathbf{o}$ being generated at state q_i
$\mathcal{N}(\mathbf{O};\boldsymbol{\mu},\boldsymbol{\sigma})$	A multivariate Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\sigma}$
$\mathbf{W}_{ij}$	Weight matrix of a deep neural network from neuron i to j
$\mathbf{v}_l$	Input vector of the l^{th} layer of a deep neural network
$\mathbf{b}_l$	Bias vector of the l^{th} layer of a deep neural network
$\mathbf{S}_{c,l}$	A speech dictionary of class c and length l
$\mathbf{N}_l$	A noise dictionary of length l
$\mathbf{A}_{c,l}$	A combined dictionary of class c and length l
$\mathbf{x}_{c,l}$	Non-negative weight vector of class c and length l
$\mathbf{y}_l$	An observation vector of length l
$\mathbf{Y}_l$	An observation matrix of length l
$\mathbf{D}$	Warping matrix
$\boldsymbol{\Lambda}$	Sparsity weight vector
$\odot$	Element-wise multiplication
$\oslash$	Element-wise division
$.[\,]$	Element-wise exponentiation
$d(y,\hat{y})$	Distance/divergence between y and $\hat{y}$
$\mathbf{1}$	A vector/matrix with all elements equal to unity

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Automatic Speech Recognition

Contemporary automatic speech recognition (ASR) research applies various statistical and data-driven pattern recognition approaches to speech signals with an eventual goal of a viable human-machine communication. Since the early attempts in 1950s to build an ASR system that can recognize a small number of words, e.g. digits, numerous approaches have been developed to achieve this task. Thanks to the enormous increase in computational power in recent years, it is feasible to incorporate many of these approaches in devices used on a daily basis such as mobile phones, tablets and computers. However, the accuracy and reliability of these systems are still much lower compared to the human speech recognition performance, creating the expectation that the current state of the art can be improved. Therefore, ASR remains to be a very progressive and active research field aiming to close the gap between the speech understanding by humans and computers.

The fundamental architecture of a generic ASR system is shown in Figure 1.1. The main blocks of an automatic speech recognizer are

- a feature extractor
- a resource repository containing an acoustic model, a language model and other resources such as a lexicon containing the phonetic information of the target language
- search space generator and decoder

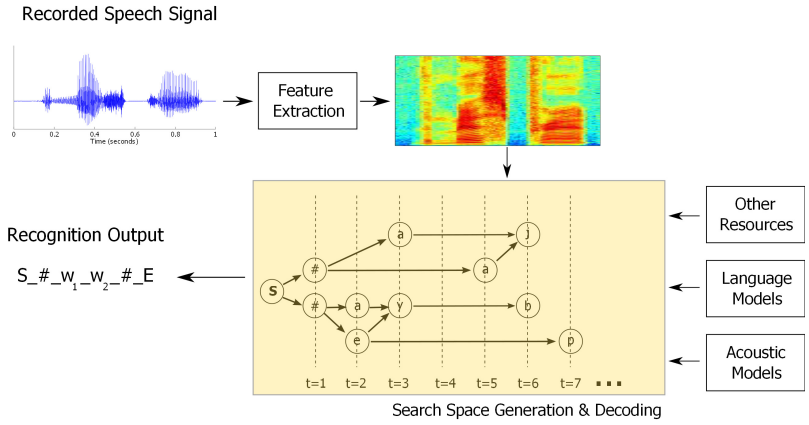


Figure 1.1: Fundamental architecture of an ASR system

Before detailing each block, we discuss how each block given in the figure interacts with the others. In a real-time ASR system, the unknown speech signals are recorded by an acoustic-to-electric transducer at a fixed sampling rate, e.g. 8k samples per second for telephony speech or 44.1k samples per second for sources such as an audio CD. This time-domain signal is processed by the feature extractor in blocks to be able to represent acoustic signals as a sequence of *feature vectors* that can be modeled with probability distribution functions such as Gaussian mixture models (GMM) or by a neural network (NN). These feature vector sequences can also be matched with reference feature vectors, i.e. templates or exemplars, with a known label to identify the label of the unknown speech segments. These statistical and data-driven models for interpreting the unknown speech segments are called the *acoustic models* and they are obtained by learning a sub-model for every speech unit that is expected to appear in the target speech. The most common choices of a speech unit are words for the recognition tasks with small vocabulary and phones for the tasks with large vocabulary.

Based on the acoustic model, the recognizer creates a search space consisting of the most probable phone/word strings (hypotheses) throughout the time. For large vocabulary recognition tasks, the recognizer also takes a language model into consideration during the search process in order to reduce the likelihood of the strings that are very unlikely to appear in the target language. The decoding algorithm identifies the single most likely hypothesis in the search space in an efficient way. Alternatively, it can create a lattice as a compact representation of the set of most likely hypotheses. In the following sections, we briefly describe these blocks with a focus on the feature extraction and acoustic

modeling which are relevant in the scope of this thesis. More information about the language models and decoding blocks can be found in [82].

1.1.1 Feature extraction

The speech samples recorded by a microphone are processed by a pre-emphasis filter to amplify the higher frequencies against the attenuation due to the lip radiation. The pre-emphasized signal is segmented into overlapping frames mostly containing a fixed number of speech samples. A typical frame duration and a shift between two frames vary between 25-30 ms and 5-10 ms respectively. Then, windowing is applied to each frame aiming to remove the spurious effects at the frame boundaries and the windowed signal is transformed into the frequency domain by applying the short-time Fourier transform (STFT). The magnitude of the complex STFT coefficients is referred to as the full-resolution spectral representation of the input speech, also called a *spectrogram*. The final speech features are extracted based on the spectrogram and it is common practice to employ several dimension reduction and decorrelation steps to remove the redundancy in the STFT features. These latter steps can also aim to project the STFT features to another feature space where the different acoustic units can be better discriminated.

In the scope of this thesis, the STFT coefficients are processed through filterbanks with an overlapping triangular-shaped frequency response to reduce the feature dimension and remove some of the speaker-dependent information such as pitch. This step is also motivated by the frequency dispersion performed by the basilar membrane in the human ear. The weighted sum of the STFT coefficients using the triangular filterbanks mimics the limited spatial resolution of the tonotopical coding by the basilar membrane. When the position and width of the filterbanks adhere to the mel frequency scale, the resulting features are called mel-scaled spectral features. It is also common to compress the magnitude range of mel-scaled spectral features by taking the logarithm and applying the discrete cosine transform (DCT) to decorrelate the features, resulting in the well-known mel-frequency cepstral coefficients (MFCC) [27]. Additionally, cepstral mean normalization (CMN) [7] is often applied to remove the effect of linear filtering (with a short impulse response). The first- and second-order time differences (often called derivatives) are concatenated to capture the inter-frame variations. Instead of the DCT, discriminatively trained transformations are also applied on the spectrogram, providing improved recognition performance. One example is the mutual information discriminant analysis (MIDA) [31] training which learns a transformation based on the mutual information criterion to discover an optimal feature subspace using the training data.

The extracted speech features $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T]$ are used in the acoustic model training and testing, i.e. recognizing unknown utterances by comparing them with the learned models of each speech unit. The choice of the speech features highly depends on the adopted acoustic modeling scheme. The techniques with non-negativity and additivity requirements commonly use the mel-scaled spectral features, while others benefit from the improved recognition accuracy of the discriminatively trained features.

1.1.2 Acoustic modeling

The definition of the acoustic modeling becomes evident when the Bayesian formulation of the speech recognition task is considered. An ASR system assigns the probability $P(\mathbf{W}|\mathbf{O})$, i.e. all possible word sequences $\mathbf{W} = [w_1, w_2, ..., w_L]$ given the speech features $\mathbf{O}$ representing the target speech signal. The word sequence $\hat{\mathbf{W}}$ with the highest probability is the recognizer output,

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{O}) \tag{1.1}$$

The probability $P(\mathbf{W}|\mathbf{O})$ can be decomposed into two parts applying the Bayes' rule as below,

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \frac{P(\mathbf{O}|\mathbf{W})P(\mathbf{W})}{P(\mathbf{O})} \tag{1.2}$$

$P(\mathbf{O})$ can be omitted as it does not depend on $\mathbf{W}$. This results in

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{O}|\mathbf{W})P(\mathbf{W}) \tag{1.3}$$

which formally defines the knowledge sources that are used in the ASR systems. $P(\mathbf{O}|\mathbf{W})$ is the *likelihood* of the observed feature sequence $\mathbf{O}$ given the word sequence $\mathbf{W}$. This probability is associated with the acoustic model. The acoustic model captures the information about the acoustic component of the speech signal aiming to classify different acoustic units accurately. Specifically, acoustic models include the reference representations of the speech units that are expected to be observed in a recognition task and they are used to assign a probability of an observed feature sequence being a phone/word sequence. We assume that the recognizer performs word recognition throughout this thesis. These words are described as a sequence of the speech units, e.g. phones, in a previously defined lexicon and each speech unit has an acoustic model based on the techniques/models described below.

$P(\mathbf{W})$ is the prior probability of the word sequence $\mathbf{W}$ and this probability is provided by the language model which is trained on a large written corpus of

the target language. Hence, the language model only depends on the target language unlike the acoustic model.

The main challenges of acoustic modeling are intra- and inter-speaker variability and adverse environmental conditions. The undesired signals that lower the intelligibility of the target speech signal are called the *background noise*. The recognition accuracy of the ASR systems reduces considerably when an acoustic model trained on noise-free or clean speech is used for recognizing speech signals degraded with background noise. This is due to the mismatch between the training and testing conditions. This mismatch has worse consequences in case of non-stationary noise and numerous approaches have been proposed to cure this problem in the literature. More details about these techniques are given in Section 1.2.

The following sections briefly summarize the most prominent statistical and data-driven models that have been used for acoustic modeling in the past. One of the oldest pattern recognition approaches that has been applied to the ASR problem is based on template or exemplar matching. Template matching is performed by measuring the similarity between the test frame sequences and the labeled training frame sequences. For this purpose, the dynamic time warping (DTW) algorithm has been adopted to increase the robustness against duration variation between the training and test utterances. In the earlier years of ASR, the computational load required to perform this comparison for a large vocabulary was a formidable task considering the available computational power. This limitation of exemplar matching shifted the interest more towards statistical approaches with compact representations. The hidden Markov models (HMM) with GMMs became the standard acoustic modeling tools for three decades due to their generalization capabilities and the development of efficient training and recognition algorithms.

Sparse representations (SR) of speech is another influential data-driven recognition approach which performs acoustic modeling based on fixed-length training frame sequences with frame-level labels. SR-based techniques have contributed to the noise robustness of automatic speech recognizers providing large improvements in the recognition accuracy under noisy conditions. Finally, the early interest in adopting artificial neural networks (ANN) in ASR systems became computationally feasible only in the last decade and ANNs with multiple layers, namely deep neural networks (DNN), are recently receiving great attention and replacing the GMMs in HMM-based recognition systems as they provide the best recognition results on large vocabulary ASR tasks as well as in many other pattern recognition tasks.

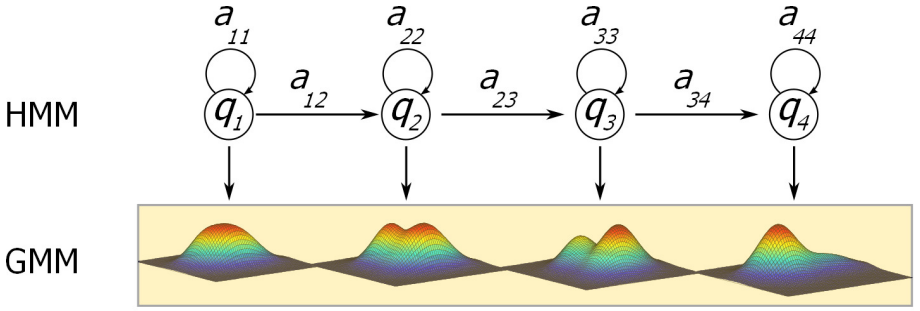


Figure 1.2: HMM models with GMM-based emission probabilities

Statistical acoustic models

1) *GMM-HMM*: The speech feature stream can be considered as a sequence of the instantaneous spectral representations of speech evolving through time. Figure 1.2 demonstrates how speech features can be generated using a left-to-right HMM with a continuous emission (output) distribution. The generation of various time series such as speech features can be modeled with an HMM which comprises sub-HMM models for every speech unit. These sub-HMMs learn a frame level representation of the target speech unit in two layers. The first layer is a first-order Markov chain containing several hidden states $\mathbf{Q} = q_0, q_1, ..., q_I$. This Markov chain is specified with a initial state probability distribution Φ_i=$P(q_0 = i)$ and a transition matrix $\mathbf{A}_{ij} = P(q_{t+1} = j|q_t = i)$. The first-order Markovian assumption implies that the probability $P(q_t|q_{t-1}, q_{t-2}, ..., q_1) = P(q_t|q_{t-1})$, i.e. being in state q_t at time t depends only on the previous state q_{t-1}. There are two types of transitions available in the left-to-right HMM structure given in Figure 1.2. The first type is a self-loop in which the transition starts and ends at the same state. The self-loop probabilities are given on the main diagonal of the transition matrix $\mathbf{A}$. This type of transition aims to handle the temporal variation in speech, e.g. for slower utterances more self-loops are picked and vice versa. The second type of transition is moving to the next state which mostly implies a more drastic spectral variation between the current and next frame compared to a self-loop.

The second layer is a discrete or continuous emission probability distribution, each belonging to a state q_i, to assign the emission probability $b_i(\mathbf{o}_t)$, i.e. the probability of the observed frame $\mathbf{o}_t$ being generated at state q_i. In case of a discrete density HMM (DDHMM), the observed frame sequences are clustered and quantized so that each frame is marked with the cluster index. Conventionally, continuous density HMMs (CDHMMs) are used with an emission

probability distribution in the form of a GMM which can be expressed as

$$P_i(\mathbf{O}) = \sum_{k=1}^{K} \mathbf{w}_{ik}\, \mathcal{N}(\mathbf{O}; \boldsymbol{\mu}_{ik}, \boldsymbol{\sigma}_{ik}) \tag{1.4}$$

where $\mathcal{N}(\mathbf{O}; \boldsymbol{\mu}, \boldsymbol{\sigma})$ is a multivariate Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\sigma}$. The indices i and k mark the state and the Gaussian distribution (mixture component) index and $\mathbf{w}_{ik}$ is the contribution of the k^{th} Gaussian to the emission probability of the state q_i. K is the total number of Gaussian distributions used in the GMM. For the sake of completeness, the expression of a G-variate Gaussian distribution $\mathcal{N}(\mathbf{O}; \boldsymbol{\mu}, \boldsymbol{\sigma})$ with a diagonal $\boldsymbol{\sigma}$, which is often used in ASR systems to reduce the computational complexity, is given below.

$$\mathcal{N}(\mathbf{O}; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{\sqrt{(2\pi)^G \prod_h \sigma_{hh}^2}} \exp\left(-\sum_g \frac{(O_g - \mu_g)^2}{2\sigma_{gg}^2} \right). \tag{1.5}$$

Training an HMM for ASR requires labeled training data so that the emission probability distributions for each speech unit can be trained accurately by applying an HMM parameter estimation technique such as the expectation maximization (EM) algorithm.

Applying the EM algorithm, the transition matrix $\mathbf{A}$, the initial state probabilities π and the emission probability distributions b_i are learned. We refer the reader to [133] for further details of the HMM training procedure. Once the HMMs representing each speech unit are trained, it is possible to compute the probability of following a particular state sequence $\mathbf{Q}$ given an observed speech feature sequence $\mathbf{O}$, $P(\mathbf{Q}|\mathbf{O})$.

$$\hat{\mathbf{Q}} = \arg\max_{\mathbf{Q}} P(\mathbf{Q}|\mathbf{O}) \tag{1.6}$$

2) *DNN-HMM*: Recent efforts have demonstrated that replacing GMMs with DNNs in an HMM system yields large improvements in the recognition accuracy [74, 188]. The idea of combining NN with HMM dates back to 1980s [165]. However, limitations on the computational power and the amount of data available for training the neural networks postponed the emergence of NN-HMM systems until the 2010s. This breakthrough was further helped by efficient pretraining algorithms [75] that avoid local extrema when estimating the network weights. Currently, context-dependent (CD)-DNN-HMM ASR systems [26], as depicted in Figure 1.3, are considered to be the state-of-the-art and they are becoming the conventional back-end for multi-stream recognition

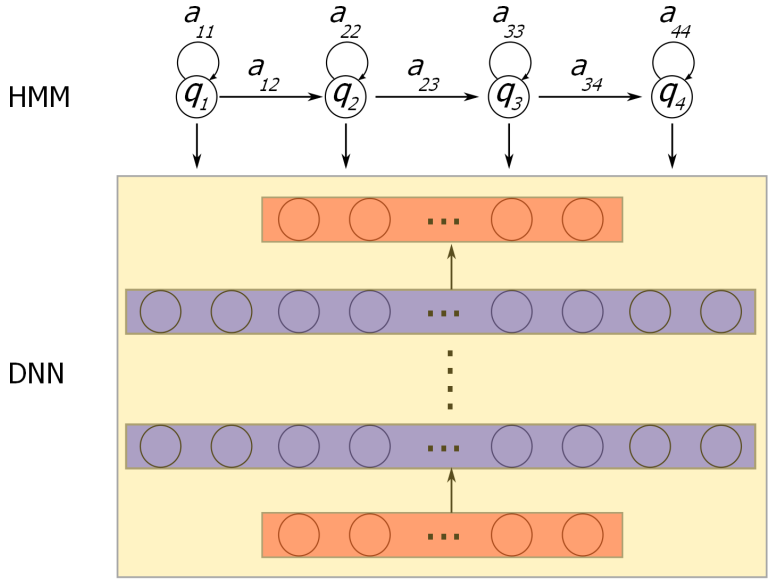


Figure 1.3: HMM models with DNN-based emission probabilities

systems. These systems are based on a traditional multilayer perceptron with an input layer, several hidden layers and an output layer.

We briefly summarize the mathematical background of CD-DNN-HMM systems starting with a single neuron structure and extending it to the complete DNN architecture. A single artificial neuron, which is the basic element of the DNN structure, gets N input values $\mathbf{v} = [v_0, v_1, ..., v_N]$ with weights $\mathbf{w} = [w_0, w_1, ..., w_N]$ and a bias value b, processes all input values to obtain the summation z and returns y as the output of a non-linear function $f(z)$,

$$y = f(z) = f(\mathbf{w}^T\mathbf{v} + b) \tag{1.7}$$

To extend the model from a single artificial neuron to a layer of M neurons, the weight vector $\mathbf{w}$ is converted into a weight matrix

$$\mathbf{W} = \begin{pmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,N} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{M,0} & w_{M,1} & \cdots & w_{M,N} \end{pmatrix}. \tag{1.8}$$

For a DNN with L hidden layers, the output of the $(l-1)^{\text{th}}$ layer with M_{l-1} neurons is the input of the l^{th} layer with M_l neurons which is formulated as

$$\mathbf{v}_l = f(\mathbf{W}_l \mathbf{v}_{l-1} + \mathbf{b_l}) \tag{1.9}$$

where the dimensions of $\mathbf{v}_l$, $\mathbf{W}_l$, $\mathbf{v}_{l-1}$ and $\mathbf{b_l}$ are $(M_l \times 1)$, $(M_l \times M_{l-1})$, $(M_{l-1} \times 1)$ and $(M_l \times 1)$ respectively. M_0 is the number of neurons in the input layer which is equal to the dimension of the speech features. The non-linear activation function f maps an M_{l-1} vector to an M_{l-1} vector and the most popular choices for f are the sigmoid function, hyperbolic tangent function and rectified linear units. The output layer has to be handled more carefully depending on the application. Assuming that the $L+1^{\text{th}}$ layer is the output layer, the activation function applied at the output layer is the softmax function in order to get output values in the range $[0, 1]$ for the HMM state posterior probabilities,

$$\mathbf{v}_{L+1} = P(q_i|\mathbf{o}) = \frac{e^{z_L^i}}{\sum_{m=1}^{M_{L+1}} e^{z_L^m}} \tag{1.10}$$

where M_{L+1} is equal to the number of HMM states.

The training of DNN-HMM systems is summarized in [26, 188] and achieved in three main stages. Firstly, a GMM-HMM setup is trained to obtain the structure of the DNN-HMM model, initial HMM transition probabilities and training labels of the DNNs. Then, the pretraining algorithm is applied to obtain a robust initialization for the DNN model. Finally, the back-propagation algorithm [69] is applied to train the DNN that will be used as the emission distribution of the HMM states. However, this procedure is not standardized yet, as alternative training techniques for DNN-HMM recognition systems are still intensively researched [35, 140, 148, 190, 191].

Data-driven acoustic models

1) *Exemplar Matching*: Exemplar or template matching is a rather straightforward and intuitive approach to tackle any pattern recognition problem in which the unknown feature vectors are compared with some reference patterns of every available class. These reference patterns are often called *templates* or *exemplars* in the literature. The acoustic model in an exemplar matching-based system contains a large collection of exemplars from each speech unit and this collection is expected to handle any kind of variation in the speech signals. Applying time warping, the exemplar matching scores become more robust to temporal variations. Therefore, it is common practice in these systems to apply DTW with certain conditions to be able to find the optimal match between two

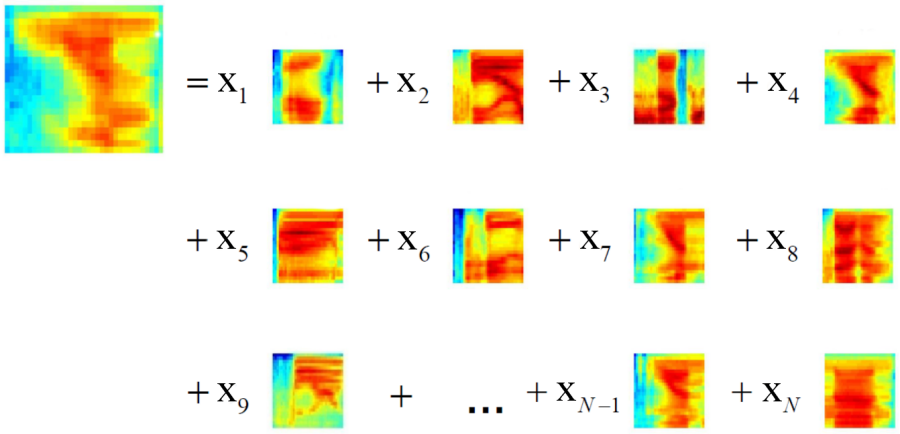


Figure 1.4: Illustration of Exemplar-based Sparse Representations of Speech

sequences of different duration. In [145], the authors define a warping function to restrict the unrealistic mappings between the time indices of the sequences by proposing an adjustment window. Hence, the warping is performed only within a window limiting the amount of stalling and skipping frames. More conditions were also set to define the behavior at the first and last frames and to avoid moving backwards in time. The recognition is performed by calculating the acoustic scores between the test utterance and exemplars taking these conditions into account. In the final phase, the decoder searches for the optimal exemplar sequence with *the minimum reconstruction error*.

This approach had to be abandoned in the early days of ASR, as the computational and memory requirements were too large for the computers used some decades ago. The computational bottleneck is the acoustic score calculation between every exemplar and the sub-segments of the target utterance and the search for the best matching exemplar sequence in a huge network of possible hypotheses. Thanks to the enormous leap in the computational power and development of fast exemplar matching algorithms, exemplar matching-based ASR techniques became popular again recently [2, 29]. However, application of the exemplar matching on large vocabulary recognition tasks is still very computationally demanding even using the most efficient exemplar selection and matching techniques on powerful processors with multiple cores.

2) *Sparse Representations*: Another exemplar-based recognition scheme, namely sparse representations, uses fixed-length training frame sequences with frame-level labels which are also called *atoms* to linearly approximate test utterances

as a linear combination of a few atoms as shown in Figure 1.4. The atoms are organized in a *dictionary* and noise robust recognition can be achieved by including atoms that contain background noise only. Hence, the SR-based system actually performs source separation on the noisy speech features by finding a few speech and noise atoms that can explain the speech and noise component in the noisy mixture. A weighted sum of these exemplars is obtained by solving a convex optimization problem with a cost function including the reconstruction error and a sparsity inducing term to avoid overfitting. The non-negative exemplar weights or activations are found by applying a multiplicative update rule that minimizes the cost function.

The mathematical formulation of the exemplar-based sparse representations of speech is as follows. Mel-scaled magnitude spectral features representing speech and noise exemplars of size $D \times L$, where D is the number of mel bands and L is the number of frames, are extracted from the training data. Each speech and noise exemplar is reshaped into a vector and stacked in the column of the speech dictionary $\mathbf{S}$ and noise dictionary $\mathbf{N}$ respectively. To be able to model noisy speech, the speech and noise dictionaries are concatenated to form the combined dictionary $\mathbf{A} = [\mathbf{S}\ \mathbf{N}]$ of size $D \cdot L \times N$ where N is the number of total speech and noise exemplars. Fixed-length segments of L frames are also extracted from a noisy utterance of T frames by applying a sliding window approach [50] and the noisy segments are reshaped into a vector to form observation matrix $\mathbf{Y} = [\mathbf{y}^1, \mathbf{y}^2, ..., \mathbf{y}^l, ..., \mathbf{y}^{T-L+1}]$ of size $(D \cdot L) \times (T - L + 1)$. An observation vector $\mathbf{y}^l$ is expressed as a linear combination of the exemplars in the combined dictionary,

$$\mathbf{y}^l \approx \sum_{n=1}^{N} \mathbf{a}_n x_n = \mathbf{A}\mathbf{x}. \tag{1.11}$$

Here, $\mathbf{x}$ is the N-dimensional non-negative exemplar weight vector. The exemplar weights are obtained by minimizing the cost function

$$d(\mathbf{y}^l, \mathbf{A}\mathbf{x}) + \sum_{n=1}^{N} x_n \lambda_n \tag{1.12}$$

where $\mathbf{\Lambda}$ is the N-dimensional non-negative vector. The first term of the cost function is the reconstruction error between the observation vector and its approximation. The second term enforces sparsity by penalizing the nonzero elements of the exemplar weight vector x. The amount of sparsity can be adjusted by assigning different values to the elements of $\mathbf{\Lambda}$. The most commonly used divergence measure used in this context is the generalized Kullback-Leibler divergence (KLD) as it provided impressive source separation, enhancement and noise robust recognition performance on speech signals. The generalized

KLD is defined as

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k. \tag{1.13}$$

The aforementioned cost function is minimized by iteratively applying the multiplicative update rule

$$\mathbf{x} \leftarrow \mathbf{x} \odot (\mathbf{A}^T(\mathbf{y}_l \oslash (\mathbf{A}\mathbf{x}))) \oslash (\mathbf{A}^T\mathbf{1} + \mathbf{\Lambda}) \tag{1.14}$$

with $\odot$ and $\oslash$ denoting element-wise multiplication and division respectively. $\mathbf{1}$ is a $(D \cdot l)$-dimensional vector with all elements equal to unity.

The recognition can be achieved after obtaining the exemplar weights by adopting several different techniques at the back-end. The first technique, *sparse classification*, infers state likelihood estimates for an HMM system and performs a modified Viterbi decoding to find the most likely state sequence. The *feature enhancement* approach reconstructs the speech component using the speech exemplars and their weights and recognizes the enhanced features using the conventional GMM-HMM recognizer trained either on original or enhanced training data. The latter acoustic models are referred to as the retrained acoustic models. Finally, *sparse imputation* has been proposed to estimate the reliable time-frequency cells from the reconstructed speech and noise components. The recognition is performed only based on the reliable regions of the spectral features.

1.1.3 Language modeling

The language model assigns a probability to word sequences, $P(\mathbf{W})$, based on their likelihood to appear in the target language. By using a language model, the speech recognizer also imposes the linguistic information during the search space creation and decoding. This results in a recognition output that is found not only based on the acoustic information, but also the grammar of the target language. There are plenty of deterministic and probabilistic language models proposed in the literature [82]. The deterministic language models such as context-free grammars are mostly used for small vocabulary recognition tasks. For large vocabulary continuous speech recognition (LVCSR), the probabilistic language models are adopted, most commonly in the form of an N-gram model. These models are trained on very large written corpora of the target language to infer the some statistics of the target language. According to these statistics, they assign a probability to every possible word based on the context information. In other words, these models aim to predict the following word based on the context, i.e. the preceding words.

For a word sequence $\mathbf{W} = [w_1, w_2, ..., w_L]$, the probability of observing $\mathbf{W}$ is given as

$$P(\mathbf{W}) = \prod_{l=1}^{L} P(w_l | w_1, w_2, ..., w_{l-1}). \tag{1.15}$$

An N-gram model concentrates only on the previous *N-1* words. Due to this assumption, $P(\mathbf{W})$ is rewritten using a limited context compared to Equation 1.15,

$$P(\mathbf{W}) \approx \prod_{l=1}^{L} P(w_l | w_{l-N+1}, w_{l-N+2}, ..., w_{l-1}). \tag{1.16}$$

In real world applications, N is typically set to $N \leq 3$ to limit the complexity. The maximum likelihood estimation of the N-gram probabilities is obtained simply by normalizing the occurrence counts of each context appearing in the written corpus,

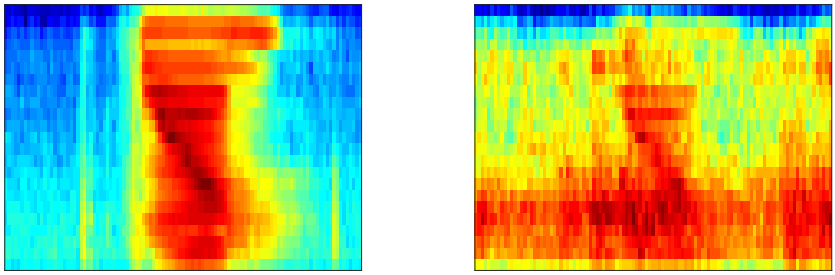
$$P(w_l | w_{l-N+1}, w_{l-N+2}, ..., w_{l-1}) = \frac{Cnt(w_{l-N+1}, w_{l-N+2}, ..., w_{l-1}, w_l)}{Cnt(w_{l-N+1}, w_{l-N+2}, ..., w_{l-1})}. \tag{1.17}$$

However, the probabilities obtained for many contexts are either equal to zero or unreliable due to lack of occurrences in the corpus. Several back-off and smoothing techniques have been proposed to cope with these problems [18,93,99].

1.1.4 Decoding

The decoding of the speech signal involves finding the most likely word sequence considering all the available resources. Various search techniques have been proposed to find the recognition output in an efficient manner, e.g. in [8, 82, 102, 124, 131]. One popular way of implementing this kind of search is to combine the scores of different models by representing each resource in the form of a finite state transducer (FST) and combining these FSTs to find the ultimate network of possible hypotheses that are allowed by all these resources [66, 120]. Some pruning can also be applied to keep the network of possible hypotheses at a tractable level. The composed FST is searched for the word sequence W that has the highest combined score which is formulated in Equation 1.3. The most likely state sequence $\hat{\mathbf{Q}}$ that best explains the observed feature sequence $\mathbf{O}$ can be found by applying the Viterbi algorithm [43, 177] and the speech units associated with the most likely states yield the recognizer output.

For exemplar matching systems, the search problem can be visualized as a three-dimensional grid search over grid points (x, y, z) which are defined by the time frame index x of the observation, time frame index y of the exemplars and the



(a) clean (b) noisy

Figure 1.5: Comparison of clean and noise speech features

associated speech unit (class) index z [128]. The exemplar sequence yielding the minimum reconstruction error is obtained by applying dynamic programming and the labels of exemplars in this sequence constitute the recognition output.

1.2 Noise Robustness in ASR systems

ASR systems, which are trained under quiet conditions and perform well under similar conditions, provide worse performance when recognizing the target utterances corrupted by background noise. This mismatch between the training and testing conditions has a serious impact on the recognition accuracy due to the severe variations in the spectrotemporal structure of the target speech. As a result, the acoustic models trained on the speech features obtained under noise-free conditions cannot model the noisy speech features. The difference between the mel-scaled magnitude spectral features of clean and noisy speech is clearly visible in Figure 1.5. The most detrimental effect is due to the fact that noise power spreads over the spectrogram blurring the formant structure and power distribution among frequencies that are intrinsic to clean speech signals.

Based on how it corrupts the clean speech signal in time domain, the background noise can be classified as additive or convolutional noise. Robustness against both kinds of background noise is investigated in this thesis. A special case of convolutional noise, where the room impulse response is much larger than the frame length, is referred to as reverberation which is defined as the persistence of the speech signal in an enclosed acoustic space due to the echoes reflecting from the boundaries. A reverberated speech signal is modeled as the convolution

of the clean speech signal with the room impulse response modeling the decay in the acoustic energy.

The severity of the background noise is commonly quantified by the ratio of the speech signal power and background noise power on a logarithmic scale. This measure is called the signal-to-noise ratio (SNR) and the recognition performance of the proposed recognition scheme will be provided at several SNR values throughout the thesis.

In the rest of the section, the literature on noise robust automatic speech recognition techniques will be revised and the novel approach that is described in this thesis will be introduced. Due to the extensive amount of research efforts in this field, we are only able to touch upon the most prominent noise robustness techniques in the next section. A more elaborate treatment is available in a recently published review paper on noise robust ASR systems [107]. The approaches proposed for improved noise robustness of the ASR systems can be classified into five main categories which are listed and briefly summarized below.

- **Multicondition training**: In this approach, acoustic model training is not only performed on clean speech data, but also on some noisy samples so that the acoustic models can handle similar noisy testing conditions [109]. In this manner, the mismatch between the training and testing conditions is aimed to be reduced. This method is quite effective under similar training and testing conditions with a drawback of a high computational cost. It is highly sensitive to the differences in the training and testing noise conditions which results in a limited discriminative power of the multicondition trained acoustic models even in case of large multicondition training data. Moreover, dynamic noise modeling is not a feasible task in this scenario as it implies the retraining of the acoustic models.

- **Robust Speech Features**: These approaches focus on noise robust speech feature extraction schemes that reduce the adverse effects of the noise on the recognition performance without modifying the acoustic models. Some of these feature extraction schemes are based on the human auditory system which is known to be quite robust against environmental noise. Others normalize various statistical properties for improved noise robustness. Several robust speech processing techniques including MFCC, perceptual linear prediction (PLP) coefficients, relative spectra (RASTA) and modulation spectrogram (MS) features are described in [12, 41, 67, 68, 71, 72, 87, 96, 98, 168, 180].

- **Feature or Speech Enhancement**: The recognition performance of the ASR systems can be considerably improved by (partly) removing

the background noise either in the signal or the feature domain. For this purpose, any speech enhancement technique can be adopted in the front-end and the enhanced signal/features can be recognized using the acoustic models trained on clean speech. Moreover, *retraining* the acoustic models is also common practice to compensate for the artifacts introduced by the denoising in the front-end, hence, better modeling of the enhanced speech. Some examples of feature and speech enhancement-based noise robust ASR systems are described in [6,32,54,73,81,104,110,121,122,156].

- **Model Compensation**: In the scope of model compensation approaches, the acoustic models are modified or extended to incorporate information about the noise sources degrading the target speech signal. Techniques such as maximum a posteriori (MAP), maximum likelihood linear regression (MLLR) and their variants can be used to adapt to new noise or environment [45,47,103,146,150]. In addition to these techniques, some other well-known approaches, namely *parallel model combination*, *HMM decomposition*, *joint compensation of additive and convolutive distortions* have been effective in noise modeling despite their high computational requirements [46,60,171].

- **Missing Data Theory and Uncertainty Techniques**: Missing data techniques estimate a time-frequency mask identifying the *reliable* and *unreliable* regions of the spectrogram evaluating the SNR values per time-frequency cell and perform recognition only based on the *reliable* regions in which the speech component dominates the background noise. There are several techniques proposed for processing the unreliable cells such as *marginalization* which simply integrates out the unreliable data and *imputation* which estimates the missing parts based on a probabilistic model conditioned on the reliable parts.

  Compared to the binary labeling of the missing data techniques, soft and continuous weighting of the frequency-time cells has been adopted in uncertainty decoding approaches. These approaches aim to integrate the uncertainty model learned in the front-end into the decoding performed by the back-end. Several examples of these techniques can be found in [24,25,39,83,100,108,136,137,155,170]

1.3 Objectives

The main objective of this thesis is to investigate the feasibility of obtaining a noise robust exemplar matching (N-REM) system by combining the two data-driven acoustic modeling approaches described in Section 1.1.2. This can be

achieved by replacing the labeled exemplars of the traditional approach with the exemplars of the sparse representations. This will create an exemplar matching technique that is intrinsically noise robust. The first exemplar technique, *exemplar matching*, performs traditional pattern matching based on a score that expresses resemblance of incoming speech and exemplars. The second technique, *sparse representations*, will model the speech as well as the noise as a linear combination of exemplars. Hence, both speech and noise are modeled as exemplars which would not be the case if we were to compensate the speech exemplars with traditional noise robustness techniques such as *parallel model compensation* [46].

There are some important differences between the two exemplar-based approaches. A first difference is that exemplar matching uses dynamic time warping to accommodate temporal differences between incoming speech and the exemplars, while the sparse representations framework does not support time warping. Still, we attempt to apply the exemplar-based noise modeling of the sparse representations approach with a compositional model to the traditional exemplar matching. This is performed by linearly combining the exemplars associated with the same speech unit and of the same length. In this new model, noise modeling can be explicitly achieved by including noise exemplars together with the speech exemplars. A first goal is to evaluate if the sparse representation formulation which combines exemplars of multiple length can work without DTW. As an alternative, we aim to develop an extension of the sparse compositional model which allows time warping in the proposed N-REM framework.

A second important difference between the two exemplar-based systems is the dissimilarity measure they use to select the best matching exemplars. Traditionally, exemplar matching uses the Euclidean distance or Mahalanobis distance on log-compressed features such as cepstra. The sparse compositional framework mostly uses the generalized Kullback-Leibler divergence on magnitude spectra. A second goal is to investigate which dissimilarity measure to use in the combined framework. For this purpose, we investigate a more flexible divergence family for comparing the noisy speech features with exemplars to find out the impact of the used metric on the recognition accuracy.

A third objective is to investigate several exemplar selection criteria to construct compact speech dictionaries for reduced computational requirements. A fourth objective is to develop an effective and efficient way of designing noise dictionaries for improved noise robustness rather than rudimentary techniques such as random noise exemplar extraction. Finally, the speech enhancement performance of the proposed scheme is explored and the results are compared with other exemplar-based approaches.

To be able to observe how well the proposed framework performs under different noise conditions, we use two popular small vocabulary databases, namely the AURORA-2 and the small vocabulary track of the second CHIME Challenge. The AURORA-2 database contains additive noise of different types to assess the recognition performance for matched and mismatched training-testing conditions. The other database contains a more challenging noise type with less stationary characteristics combined with a mild degree of reverberation.

1.4 Thesis Overview

This section gives a short overview of the chapters contained in this thesis which introduces a novel noise robust automatic speech recognition scheme by introducing noise modeling capabilities to exemplar matching-based acoustic modeling. This is achieved by combining exemplar-based sparse representations and exemplar matching. More specifically, exemplars associated with speech units that are used in exemplar matching-based acoustic modeling are used in a conventional exemplar-based sparse representations formulation. As a result of the multiple-length exemplars, the proposed recognizer uses multiple speech dictionaries, each containing exemplars associated with the same speech unit and of the same duration.

The inherent noise modeling problem of exemplar matching-based techniques originates due to the intractable task of evaluating all possible alignments of speech and noise exemplars to be able to perform source separation. In other words, it is not possible to discover the speech and noise components of a noisy mixture by comparing with individual speech and noise exemplars due to the enormous number of possible alignments. In our approach, we remedy this problem by approximating noisy speech features as a linear combination of speech and noise exemplars of all available exemplar lengths. This additivity is a reasonable approximation if signals are represented as mel-scaled magnitude spectral features. The decoding is performed based on the reconstruction errors of each dictionary similar to the traditional exemplar matching.

The initial investigation of the proposed exemplar matching system focuses on clean speech recognition by only using speech exemplars and comparing the test segments and exemplars of the same length. These experiments will establish the basics of the proposed exemplar matching using a sparse representation model. Then, we further investigate a model that can accommodate time warping in the new proposed setting and evaluate the clean speech performance of the system with time warping. Finally, various exemplar selection criteria

have been proposed for the undercomplete speech dictionaries and the decrease in the recognition accuracy with increasing pruning rate will be explored.

The noise robust exemplar matching concept will be introduced after clean speech experiments with a focus on the adaptive noise exemplar extraction technique. This adaptive noise modeling approach considerably increases the recognition performance of the proposed approach especially under severe noise conditions. In addition to this technique, we look into the recognition performance by adopting a more flexible divergence family, namely alpha-beta (AB) divergence, in place of the conventional generalized Kullback Leibler divergence. Having two parameters, the AB divergence provides improved robustness against the background noise.

In the last part of thesis, the speech enhancement performance of the proposed framework will be investigated by comparing the noise suppression performance with other baseline enhancement systems. In addition to these experiments, the novel speech enhancement system is employed in the front-end of a conventional GMM-HMM recognition system to evaluate the impact of the front-end denoising on the recognition performance. A general discussion, a list of the original contributions and directions for future research conclude this thesis. A brief summary of each chapter is provided below.

- **Chapter 2**: The procedure to combine exemplar matching and exemplar-based sparse representation approaches is described in this chapter. The recognition performance of the combined system is compared with a simple exemplar matching recognizer which uses discriminatively trained features and classifies the test segments as the label of the closest exemplar with respect to the Euclidean distance.

  This chapter is adapted from: Emre Yılmaz, Dirk Van Compernolle and Hugo Van hamme, "*Combining Exemplar-based Matching and Exemplar-based Sparse Representations of Speech*", In Symposium on Machine Learning in Speech and Language Processing (MLSLP), Portland, USA, September 2012.

- **Chapter 3**: This chapter describes a new sparse representation model that allows time warping as an extension to the combined system. Even though the new model with time warping has an increased computational complexity, the initial results have shown the feasibility of the approach.

  This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Embedding Time Warping in Exemplar-based Sparse Representations of Speech*", In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 8076-8080, Vancouver, Canada, May 2013.

- **Chapter 4**: Numerous speech exemplar selection criteria have been proposed and the pruned dictionaries are used for clean speech recognition to assess the quality of the chosen exemplars in terms of the recognition performance. The results have shown that up to 70% of the exemplars can be discarded without a significant loss in the recognition accuracy.

  This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Exemplar Selection Techniques for Sparse Representations of Speech Using Multiple Dictionaries*", In 21st European Signal Processing Conference (EUSIPCO), pages 1-5, Marrakesh, Morocco, Sept. 2013.

- **Chapter 5**: Noise robustness of the proposed model with an adaptive noise dictionary design technique is thoroughly investigated on two popular databases. The results demonstrate the effectiveness of the noise robust exemplar matching framework on small vocabulary ASR tasks.

  This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Noise Robust Exemplar Matching Using Sparse Representations of Speech*", IEEE/ACM Transactions on Audio, Speech, and Language Processing, volume 22, No. 8, pages 1306-1319, Aug. 2014.

- **Chapter 6**: The adaptive noise modeling scheme has been investigated in detail by looking for the optimal design parameters to find a trade-off between the computational complexity and the noise robustness. Using the optimal parameters boosts the recognition performance without a noticeable increase in the computational burden.

  This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Adaptive Noise Dictionary Design for Noise Robust Exemplar Matching of Speech*", Submitted to EUSIPCO 2015.

- **Chapter 7**: The alpha-beta divergence has been used to learn the noisy speech approximation and calculate the reconstruction errors used at the back-end. Adjusting the parameters yields improved noise robustness and an elaborate discussion is given on the divergence parameter choice.

  This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Noise Robust Exemplar Matching with Alpha-Beta Divergence*", Submitted to Speech Communication, 2015.

- **Chapter 8**: The speech dictionary design issue is revisited taking the novel divergence measure into account. The best performing criterion of the previous work and a novel k-medoids technique using the alpha-beta divergence is compared with random selection and a baseline exemplar technique performing well for previous sparse representations techniques.

This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Data Selection for Noise Robust Exemplar Matching*", Submitted to INTERSPEECH 2015.

- **Chapter 9**: The noise robust exemplar matching framework is used as a single-channel speech enhancement technique and the enhancement quality is compared with other speech enhancement techniques. The proposed enhancement system provides superior performance with respect to several quality metrics.

  This chapter is adapted from: Emre Yılmaz, Deepak Baby and Hugo Van hamme, "*Noise Robust Exemplar Matching with Coupled Dictionaries for Single-Channel Speech Enhancement*", Submitted to EUSIPCO 2015.

- **Chapter 10**: The final chapter discusses the application of the novel single-channel speech enhancement technique to automatic speech recognition. In practice, the enhancement system is used at the front-end to reduce the amount of degradation due to the background noise. The recognition results of this speech enhancement-based technique are compared with other state-of-the-art noise robust systems.

  This chapter is adapted from: Emre Yılmaz, Deepak Baby and Hugo Van hamme, "*Noise Robust Exemplar Matching for Speech Enhancement: Applications to Automatic Speech Recognition*", Submitted to INTERSPEECH 2015.

- **Chapter 11**: This chapter concludes the thesis by listing the original contributions and directions for future research.

Chapter 2

Combining Exemplar Matching and Sparse Representations

In this chapter, we compare two different frameworks for exemplar-based speech recognition and propose a combined system that approximates the input speech as a linear combination of exemplars of variable length. This approach allows us not only to use multiple length exemplars, each representing a certain speech unit, but also to jointly approximate input speech segments using several exemplars. While such an approach is able to model noisy speech, it also enforces a feature representation in which additivity of the effect of signal sources holds. This is observed to limit the recognition accuracy compared to e.g. discriminatively trained representations. We investigate the system performance starting from a baseline single-neighbor exemplar matching system using discriminative features to the proposed combined system to identify the main reasons of recognition errors. Even though the proposed approach has a lower recognition accuracy than the baseline, it significantly outperforms the intermediate systems using comparable features.

This chapter is adapted from: Emre Yılmaz, Dirk Van Compernolle and Hugo Van hamme, "*Combining Exemplar-based Matching and Exemplar-based Sparse Representations of Speech*", In Symposium on Machine Learning in Speech and Language Processing (MLSLP), Portland, USA, September 2012.

2.1 Introduction

Exemplar-based (or template-based) speech recognition recently regained popularity due to the significant increase in computational power and development of fast template matching and search algorithms [30]. Several hybrid recognition systems combining this approach with hidden Markov models (HMMs) are also proposed [1, 9]. Exemplars are labeled speech segments such as phones or syllables, possibly of different length, that have occurred in the training data and they are matched with the input speech segments using dynamic time warping (DTW). We refer to this approach as *exemplar matching*. This approach allows to use any choice of frame-synchronous feature vector to represent the input speech and the exemplars. For instance, in [30], motivated by a better recognition accuracy, a mutual information based discriminant analysis (MIDA [31]) is applied to log-spectral data.

One can simply classify the segment as the label of the closest exemplar, or by a voting scheme on the set of K nearest neighbors [30, 57]. Applying exemplar matching under noisy conditions creates mismatch problems similar to what is experienced with HMMs. One can resort to feature compensation methods to increase the robustness to noise [59]. Model compensation techniques would require all exemplars to be modified, which is a formidable task in the case of non-stationary noise. Since the search problem in exemplar-based recognition is a lot more involved than in HMM-based recognition, the equivalent of factorial models is also not a trivial path to walk. Finally, multi-condition training, i.e. storing noisy exemplars, will increase the number of exemplars dramatically. Furthermore, noisy exemplars can only capture a certain instance of speech and noise resulting in a limited noise modeling especially in case of non-stationary noise.

More recently, exemplar-based *sparse representations* have been used successfully for clean [48, 142] and noisy [49, 84, 162] speech recognition. This technique models input speech segments as a sparse linear combination of fixed-length exemplars. These exemplars are represented in the linear magnitude spectral domain to ensure additivity. By combining speech and noise exemplars linearly, it explicitly models the noisy speech. Because exemplars are combined linearly, they need to be of the same length, unlike in exemplar matching, and cannot model our choice of speech segments (phones, syllables, ...). The exemplars can therefore not serve directly as an acoustic model, so sparse representations have been used for speech enhancement, a model of state likelihoods (sparse classification) or to generate a mask in a missing data recognition framework.

In this chapter, we elaborate on the differences between the DTW and sparse representation exemplar techniques and propose a procedure to combine them.

This results in a basic exemplar matching recognizer having the advantage of using long exemplars of variable length in a sparse representation formulation. The main motivation is to establish a new framework that allows noise modeling for exemplar matching based recognition systems. This task involves both the selection of the appropriate representation domain of speech and the distance/divergence measure used for comparing the input speech segments with exemplars. Most exemplar matching techniques make use of state-of-the-art features with high discriminative power among the classes to lower the recognition errors [30, 57]. However, as additivity and non-negativity properties are required for linearly combining exemplars, mel-scaled magnitude and power spectra can be used to represent speech in the proposed approach. The Euclidean distance used in exemplar matching has to be replaced by e.g. the generalized Kullback-Leibler divergence. This study focuses on the price that needs to be paid in terms of the accuracy on *clean* data for these modifications. An analysis of the resulting noise robustness is the topic of other work currently under review.

The rest of the chapter is organized as follows. Section 2.2 explains exemplar matching based recognition, exemplar-based sparse representations of speech and the combined system. The experimental setup is discussed in Section 2.3. Section 2.4 presents the results. The conclusions are discussed in Section 2.5.

2.2 Exemplar-based recognition systems

2.2.1 Exemplar-matching

This technique compares the input speech segments with labeled exemplars, each representing a certain class. The exemplars are collected from a large corpus that is segmented in terms of the desired classes. The segments will have variable lengths, so the natural duration distribution of each class in the training corpus is preserved. Input speech and exemplars are represented using state-of-the-art speech features in order to maximize recognition accuracy. Recognition then consists of finding the sequence of exemplars that best matches the input subject to lexical and grammatical segment concatenation constraints. The quality of a match is measured by a metric (e.g. Euclidean distance) that expresses how well the exemplars reconstruct the data. Additional constraints are imposed. Each exemplar is tagged with meta-information such as speaker characteristics (e.g. gender, age) or prosodic information (e.g. speaking rate, position in the sentence). This information is used during decoding to penalize inconsistent exemplar sequences (e.g. mixed gender) with various concatenation costs. In the present work, only two types of concatenation costs are considered, namely

exemplar startup costs and gender costs. Exemplar startup costs penalize longer exemplar sequences and control the insertion/deletion rate. Gender costs penalize mixed gender exemplar sequences, a constraint which has been shown to improve the recognition accuracy [30]. Finally, in earlier exemplar matching work, strict matches across the time dimension were relaxed using DTW. In this work, time warping is not applied for three reasons. Firstly, it would complicate the distance calculation. Secondly, in noisy conditions, too much freedom in time warping may lead to unrealistic warping, so duration constraints are more important than in clean conditions. The same effect has been observed in HMM systems [101]. Thirdly, in the combined system described in Section 2.2.3, the linear combination of exemplars with different internal time warping will relax the requirement for strict matching along the time axis.

2.2.2 Sparse Combinations of Exemplars

The exemplar-based sparse representations approach models the input speech as a linear combination of several speech exemplars [48]. The input speech and exemplars are represented in the linear mel-scaled spectral domain in order to ensure additivity of exemplars. In this framework, exemplars are fixed-length speech segments randomly extracted from the training corpus and may be associated with more than one class. Labeling is performed probabilistically using a conventional HMM-based recognizer either at the word or state level.

Exemplars consisting of L frames are reshaped as a single column vector and collected in a single dictionary $\mathbf{S}$ of dimensionality $DL \times N$ where D is the number of frequency bands and N is the number of available exemplars. A reshaped input speech vector $\mathbf{y}_L$ of length L is expressed as a linear combination of the exemplars with non-negative weights:

$$\mathbf{y}_L \approx \sum_{m=1}^{N} x_m \mathbf{s}_m = \mathbf{S}\mathbf{x} \qquad \text{s.t.} \qquad x_m \geq 0 \tag{2.1}$$

where $\mathbf{x}$ is an N-dimensional sparse weight vector. Sparsity of the weight vector implies that the input speech is approximated by a small number of exemplars. The exemplar weights are obtained by minimizing the cost function,

$$d(\mathbf{y}_L, \mathbf{S}\mathbf{x}) + \sum_{m=1}^{N} x_m \Lambda_m \qquad \text{s.t.} \qquad x_m \geq 0 \tag{2.2}$$

where $\mathbf{\Lambda}$ is an N-dimensional vector. The first term is the divergence between the input speech vector and its approximation. A regularization term is added in order to limit the l_1-norm of the weight vector. Here, $\mathbf{\Lambda}$ controls how sparse

the resulting vector $\mathbf{x}$ is. The generalized Kullback-Leibler divergence (KLD) is used for d:

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k \tag{2.3}$$

which is commonly used in source separation problems and shown to produce better results than Euclidean distance when used with linear mel-scaled spectra [174].

The regularized convex optimization problem can be solved using various methods including LASSO and non-negative sparse coding (NSC). For NSC, the multiplicative update rule to minimize the cost function (2.2) is derived in [50] and found as

$$\mathbf{x} \leftarrow \mathbf{x} \odot (\mathbf{S}^T(\mathbf{y}_L \oslash (\mathbf{S}\mathbf{x}))) \oslash (\mathbf{S}^T\mathbf{1} + \mathbf{\Lambda}) \tag{2.4}$$

with $\odot$ and $\oslash$ denoting element-wise multiplication and division respectively. $\mathbf{1}$ is a DL-dimensional vector with all elements equal to one. Applying this update rule iteratively, the weight vector becomes sparse and the reconstruction error between the input speech vector and its approximation decreases monotonically.

In order to decode the input speech, a window of length L is slid over the input speech with a constant frame shift and the weight vector for each window is obtained. Then, using a label matrix containing the word or state based labels for each exemplar, the HMM likelihood scores are calculated. Finally, a modified Viterbi algorithm is applied to find the most likely class sequence.

2.2.3 Combined System

The combined system aims to benefit from the advantages of the two frameworks explained in the previous sections. It is an exemplar matching approach in the sense that it explains the input as the sequence of classes leading to a minimal reconstruction error, each class being represented by exemplars of variable length. The reconstruction error is however measured by the sparse combination model in the linear spectral domain, which has the advantage of easily modeling noisy speech by adding noise exemplars. The exemplars are thus organized in multiple dictionaries $\mathbf{S}_{c,l}$: one for each class c and each length l as shown in Figure 2.1. Each dictionary is of dimensionality $Dl \times N_{c,l}$ where $N_{c,l}$ is the number of available exemplars of length l and class c. Using separate dictionaries for different classes is expected to provide better classification than using a single dictionary as every input segment is guaranteed to be approximated by a combination of exemplars belonging to the same class only.

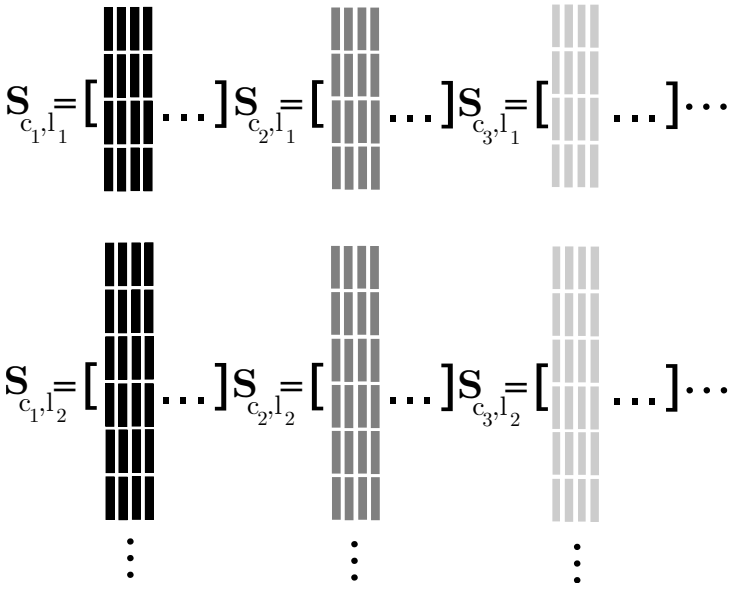


Figure 2.1: Exemplars are organized in multiple dictionaries $\mathbf{S}_{c,l}$ for each class c and each length l.

For any class c, a reshaped input speech vector $\mathbf{y}_l$ of length l is expressed as a linear combination of the exemplars with non-negative weights:

$$\mathbf{y}_l \approx \sum_{m=1}^{N_{c,l}} x_{c,l}^m \mathbf{s}_{c,l}^m = \mathbf{S}_{c,l}\mathbf{x}_{c,l} \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \tag{2.5}$$

where $\mathbf{x_{c,l}}$ is an $N_{c,l}$-dimensional sparse weight vector. The class and length dependent weight vectors are obtained by applying the multiplicative update rule in Equation (2.4) for each dictionary. The reconstruction error between a class c and an input speech segment of length l can be calculated using Equation (2.3). It satisfies the conditions to apply dynamic programming, hence the class sequence that best matches the input speech can be simply found.

The input speech is decoded similar to the exemplar matching based recognizer. Every input frame sequence of each available exemplar length is approximated as a linear combination of exemplars by iteratively applying the update formula. For each class and exemplar length, the approximation is performed separately using the dictionaries. After a certain number of iterations, the reconstruction error is calculated using Equation (2.3). As every dictionary contains exemplars with known labels, the entire input utterance is searched to find the class sequence yielding the minimum reconstruction error.

A known problem of sparse representation approaches working on magnitude spectra is that the silence exemplars are not recognized [50]. This is due to the fact that silence is well-approximated by combining speech exemplars with small weights, so all classes will score equally well. To overcome this problem, reconstruction errors for the class representing silence have to be compensated. This is achieved by reducing the reconstruction errors corresponding to silence dictionaries by a compensation factor CF depending on the voice activity value assigned to the middle frame of the corresponding input speech segment and the reconstruction error itself,

$$CF = C \cdot d(\mathbf{y}_l, \mathbf{S}_{sil,l}\mathbf{x}_{sil,l}) \cdot VAD \tag{2.6}$$

where C is a scale factor and VAD is the voice activity estimate (0 for speech, 1 for silence). The VAD value can either be obtained from an autonomous module implementing a preferred method from the vast literature on the topic, or it can be estimated using the exemplar weights $\mathbf{x}_{\mathbf{c},\mathbf{l}}$. In this work, an energy-based VAD is used. It should be noted that including the reconstruction error itself in Equation (2.6) compensates for length differences.

2.3 Experimental Setup

2.3.1 Data, Preprocessing and Features

We have conducted recognition experiments on the 4 clean test sets of the AURORA-2 database [77]. To reduce simulation time, we subsampled each test set by a factor of 4, bringing the total number of utterances to 1001. For feature extraction, a 17 channel Mel-scaled filter bank with triangular magnitude response is computed from a spectral analysis with a window length of 32 ms and a frame shift of 10 ms. The first channel is centered at 200 Hz and the last at 3030 Hz. Channel normalization of the magnitude spectrum is achieved by transforming it to the log-domain, applying mean normalization and moving back to the linear domain. The exemplar matching baseline uses MIDA features, i.e. a discriminatively trained linear transform of the mean-normalized log-power spectra and its first and second order differences (a total of $3 \times 17 = 51$ features) resulting in 32-dimensional feature vectors.

2.3.2 Exemplar Matching

The exemplars used in both the exemplar matching and in the combined system are half-digits which are extracted from the clean training set and segmented

by a conventional HMM-based system. As argued before, the design strives for long units. Full digits turned out to be too long for matching without DTW resulting in a high error rate. With half-digits the exemplars seemed to generalize sufficiently to unseen data resulting in an acceptable baseline (see below). This results in 49,354 exemplars belonging to 22 half-digit classes and 14,418 silence exemplars (in total 63,772 exemplars). The minimum and maximum exemplar lengths are 5 and 30 frames respectively. Exemplars longer than 30 frames are omitted to limit the number of dictionaries that are to be used in the further steps of the experiment.

Speech segments are classified as their single closest neighboring exemplar (1-NN). The exemplar startup and gender costs are tuned manually for maximal recognition accuracy.

2.3.3 Combined System

In the combined system, there are in total 508 dictionaries containing the same speech exemplars as in the exemplar matching baseline. However, only 1300 silence exemplars (50 exemplars for each length) are used since silence exemplars do not contribute much as discussed in Section 2.2.3. The system ends up using 50,654 exemplars in total. In the combined system, the l_2-norm of each dictionary column is set to unity, i.e. the energy of each exemplar is normalized. The same normalization is applied to the reshaped input speech vectors. The reconstruction error shows enough discrimination among classes after 50 iterations. All elements of $\mathbf{\Lambda}$ are set to 2. The scale factor C is set to 0.5. The combined system only uses exemplar startup cost which is, like for the exemplar matching system, tuned for maximal accuracy.

2.3.4 Reconstruction Error Metrics

The log-compressed features that are used in the exemplar matching baseline and the first intermediate system are compared using the Euclidean distance. All the other intermediate systems and the final system use the generalized KLD to calculate the reconstruction error.

2.4 Results and Discussion

In this section, we migrate the 1-NN exemplar matching system in several steps towards the final combined design. The steps dealing with feature representation

Table 2.1: Word error rates for the 1-NN exemplar matching based recognizer in percentages

Features	Dimension	Dis./Div. Measure	WER (%)
MIDA	32	Eucl.	1.11
MN+logPowSpec	17	Eucl.	3.36
MN+PowSpec	17	KLD	10.10
MN+MagSpec	17	KLD	4.41
PowSpec	17	KLD	10.34
MagSpec	17	KLD	4.36

Table 2.2: Word error rates for the proposed system in percentages

Features	Dimension	Dis./Div. Measure	WER (%)
PowSpec	17	KLD	7.70
MagSpec	17	KLD	2.98
l_2-N+PowSpec	17	KLD	5.14
l_2-N+MagSpec	17	KLD	2.16

and distance metric in a single nearest neighbor exemplar matching context are summarized in Table 2.1.

Based on prior design experience, we start from a design using MIDA features, channel (mean) normalization and Euclidean distance resulting in a word error rate (WER) of 1.11%. Since the sparse representation approach does not use linear transforms or derivatives, we remove this first, resulting in 3.36% WER. The second and the third lines compare the recognition accuracies obtained using mean normalized log-compressed power spectra and mean normalized linear power spectra in conjunction with the Euclidean distance and generalized KLD respectively. It can be concluded that log-compression combined with the Euclidean distance performs much better. The results in the middle and lower panels of Table 2.1 show that the generalized KLD couples much better with linear magnitude spectra and mean normalization is not effective both for magnitude and power spectra in this task.

The upper panel of Table 2.2 presents the WER obtained with the combined system using power and magnitude spectra. Compared to the lower panel of Table 2.1, there is a significant improvement on recognition accuracies in both magnitude and power spectra due to sparse combination approach. Finally, the l_2-norm of the exemplars is set to unity as described in Section 2.3.3 which boosts the recognition accuracy. Even though the best result obtained

with the proposed approach is still behind the baseline system, it significantly outperforms all the other intermediate systems with comparable features.

2.5 Conclusions

We discussed two different exemplar-based recognition schemes, namely exemplar matching and exemplar-based sparse representations, and proposed a combined system that uses multiple length exemplars to jointly approximate the input speech. Such a design can benefit from the noise model provided by the sparse representations approach while it can decode unseen speech directly in terms of exemplar identities using a reconstruction error metric. Exemplars are organized in separate dictionaries which are expected to provide better classification than using a single dictionary as every input segment is approximated by a combination of exemplars belonging to the same class only. The additivity and non-negativity requirement limits the representation domain to magnitude or power spectra. This apparently leads to lower recognition accuracy compared to discriminatively trained speech features. Moreover, the Euclidean distance, which is widely used in exemplar matching based systems, has to be replaced by the generalized KLD.

Chapter 3

Embedding Time Warping

This chapter describes a new sparse representation model for speech that allows time warping as an extension to a recently proposed sparse representations-based speech recognition system. This recognition system uses exemplars to model the acoustics which are labeled speech occurrences of different length extracted from the training data. Exemplars are organized in multiple dictionaries on the basis of their class and length. Input speech segments are approximated as a sparse linear combination of the exemplars using these dictionaries and a reconstruction error-based decoding is adopted in order to find the best matching class sequence. With the current sparse representation model using a dictionary and a weight vector to approximate an input speech segment, it is not possible to compare input speech segments with exemplars of different lengths. The goal of this work is to introduce a novel sparse representation model which allows time warping using a third matrix which linearly combines consecutive frames in order to shrink or expand the approximation. The results have shown the feasibility of the proposed sparse representation model.

This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Embedding Time Warping in Exemplar-based Sparse Representations of Speech*", In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 8076-8080, Vancouver, Canada, May 2013.

3.1 Introduction

Automatic speech recognition has been dominated by statistical acoustic modeling tools, e.g. Hidden Markov models, for several decades. The success of recently proposed speech recognition systems based on exemplar matching attracted considerable interest in exemplar-based acoustic modeling as a viable alternative [143]. These techniques use real speech data, either called exemplars or templates, to recognize unseen speech. Exemplars are labeled speech segments such as phones, syllables or words, possibly of different length, that are extracted from the training data. Each exemplar is tagged with meta-information including speaker, environmental characteristics and prosodic information. Inconsistent exemplar sequences, e.g. mixed gender exemplar sequences, can be penalized based on the tagged meta-information during recognition. An input speech segment can be classified by evaluating the labels of the closest exemplars obtained using a distance metric.

Although exemplars provide better duration and trajectory modeling compared to Hidden Markov Models, they are poorer in terms of generalizability. To cope with this shortcoming, large amounts of data are required to handle the acoustic variation among different utterances [30]. Furthermore, the acoustic distance between the input speech segments and exemplars is found using the dynamic time warping algorithm (DTW). DTW is a well-known algorithm used for matching frame sequences of different lengths in various applications such as speech recognition [2, 29, 145], image recognition [56], audio classification [132] and data mining [14].

An alternative exemplar-based recognition technique is called exemplar-based sparse representations (SR) in which the spectrogram of input speech segments is modeled as a sparse linear combination of exemplars of the same length. SR-based techniques have been successfully used for speech enhancement [54], feature extraction [142] and clean [48] and noisy [50, 84, 162] speech recognition. We have recently proposed an SR-based speech recognition system which uses exemplars of different length organized in separate dictionaries on the basis of their class and length [187]. Compared to a system using fixed-length exemplars stored in a single dictionary, using separate dictionaries for each class provides better classification as input speech segments are approximated as a linear combination of exemplars belonging to the same class only. We have also shown that this system performs reasonably well under noisy conditions in [181].

According to our knowledge, previous SR-based speech recognition systems do not embody a time warping mechanism that allows the comparison of the different-length segments. This chapter proposes a novel sparse representation model of speech that embeds time warping in the previous model consisting of a

dictionary and a weight vector. Time warping is achieved by means of a sparsely structured warping matrix that learns weights to linearly combine corresponding frequency bands in consecutive frames. The design of the warping matrix has to be handled carefully as too much flexibility in time warping may lead to unrealistic warping. Therefore, only a few successive frames should be combined to approximate an input speech frame. Moreover, sparsity regularization is imposed on the warping matrix to obtain linear combinations often dominated by a single frequency band. This constraint results in approximations that are close to one of the actual frequency bands rather than random linear combinations.

The proposed system differs from classical DTW in several aspects. One main difference is that the proposed model performs a frequency band-level warping by learning distinct weights for each frequency band in a frame, whereas classical DTW provides a frame-level mapping between the time axes. The proposed warping scheme is expected to be more robust against spectral asynchronies, i.e. channel effects in the form of frequency-dependent delays, as it is able to compensate temporal jitters depending on the number of linearly combined successive frames, e.g. when two successive frames are linearly combined, a spectral asynchrony with temporal jitter of a frame shift (typically 10 ms) can be handled. In this sense, the proposed recognizer better models human hearing which is not sensitive to spectral asynchronies up to 40 ms [3].

The rest of the chapter is organized as follows. The proposed sparse representation model allowing time warping is given in Section 3.2. Section 3.3 explains the experimental setup and implementation details. In Section 3.4, we present the recognition results and a discussion on the proposed model and its relations with classical DTW is given. The conclusions and thoughts for future work are discussed in Section 3.5.

3.2 Sparse Representation Model of Speech with Time Warping

3.2.1 Previous Model

In the sparse representation model described in [187], the input segments are modeled as a linear combination of the exemplars that are stored in the dictionaries. Each exemplar represents a certain speech unit and the duration of each speech unit in the training data is preserved resulting in exemplars of different lengths. Exemplars spanning l_e frames are reshaped into a single vector with Fl_e time-frequency cells where F is the number of frequency bands in a frame. These reshaped exemplars are stored in the columns of the dictionary

$\mathbf{S}_{c,l_e}$: one for each speech unit c and each length l_e. Each dictionary is of dimensionality $Fl_e \times N_{c,l_e}$ where N_{c,l_e} is the number of available exemplars of class c and length l_e.

The baseline model approximates a reshaped input speech vector $\mathbf{y}_{l_i}$ of length Fl_i as a linear combination of the reshaped exemplars of length Fl_e with non-negative weights for each class c:

$$\mathbf{y}_{l_i} \approx \sum_{m=1}^{N_{c,l_e}} \mathbf{s}^m_{c,l_e} x^m_{c,l_e} = \mathbf{S}_{c,l_e}\mathbf{x}_{c,l_e} \quad \text{s.t.} \quad x^m_{c,l_e} \geq 0 \tag{3.1}$$

where $l_i = l_e$ and $\mathbf{x}_{c,l_e}$ is an N_{c,l_e}-dimensional sparse weight vector. Sparsity of the weight matrix implies that the input speech is approximated by a small number of exemplars. The exemplar weights are obtained by minimizing the cost function,

$$d(\mathbf{y}_{l_i}, \mathbf{S}_{c,l_e}\mathbf{x}_{c,l_e}) + \Lambda \sum_{m=1}^{N_{c,l_e}} x^m_{c,l_e} \qquad \text{s.t.} \qquad x^m_{c,l_e} \geq 0 \tag{3.2}$$

where Λ is a scalar that controls how sparse the resulting vector $\mathbf{x}_{c,l_e}$ is. The first term is the divergence between the input speech vector and its approximation. The second term is a regularization term which penalizes the l_1-norm of the weight vector to produce a sparse solution. The generalized Kullback-Leibler divergence (KLD) is used for d:

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k \tag{3.3}$$

The regularized convex optimization problem can be solved using various methods including non-negative sparse coding (NSC). For NSC, the multiplicative update rule to minimize the cost function (3.2) is derived in [50] and is given by

$$\mathbf{x}_{c,l_e} \leftarrow \mathbf{x}_{c,l_e} \odot (\mathbf{S}^T_{c,l_e}(\mathbf{y}_{l_i} \oslash (\mathbf{S}_{c,l_e}\mathbf{x}_{c,l_e}))) \oslash (\mathbf{S}^T_{c,l_e}\mathbf{1} + \Lambda) \tag{3.4}$$

with $\odot$ and $\oslash$ denoting element-wise multiplication and division respectively. $\mathbf{1}$ is a Fl_e-dimensional vector with all elements equal to unity.

3.2.2 Proposed Model

To be able to generalize the approximation in Equation (3.1) to input speech segments of length l_i for $l_i \neq l_e$, we introduce a sparse warping matrix $\mathbf{D}_{c,l_i,l_e}$

of dimensionality $Fl_i \times Fl_e$. For the sake of conciseness, we use $\mathbf{D}$, $\mathbf{S}$, $\mathbf{x}$ and N to represent $\mathbf{D}_{c,l_i,l_e}$, $\mathbf{S}_{c,l_e}$, $\mathbf{x}_{c,l_e}$ and N_{c,l_e} respectively. This warping matrix linearly combines the successive frames to shrink or expand the approximation $\hat{\mathbf{y}}_{l_e} = \mathbf{S}\mathbf{x}$. Thus, a reshaped input speech vector $\mathbf{y}_{l_i}$ can be approximated as a linear combination of the time-frequency cells belonging to successive frames in $\hat{\mathbf{y}}_{l_e}$ for $l_i \neq l_e$,

$$\mathbf{y}_{l_i} \approx \sum_{n=1}^{Fl_e} \mathbf{d}^n y_{l_e}^n = \mathbf{D}\hat{\mathbf{y}}_{l_e} \tag{3.5}$$

where $\mathbf{d}^n$ is the n^{th} column of the warping matrix $\mathbf{D}$. Combining Equation (3.1) and (3.5), the complete model can be written as

$$\mathbf{y}_{l_i} \approx \sum_{n=1}^{Fl_e} \sum_{m=1}^{N} \mathbf{d}^n s^{n,m} x^m = \mathbf{D}\mathbf{S}\mathbf{x} \quad \text{s.t.} \quad x^m, d^{n,m} \geq 0. \tag{3.6}$$

The new cost function is comprised of three components,

$$d(\mathbf{y}_{l_i}, \mathbf{D}\mathbf{S}\mathbf{x}) + \Lambda \sum_{m=1}^{N} x^m + \beta \sum_{n=1}^{Fl_i} \sum_{m=1}^{Fl_e} d^{n,m} \quad \text{s.t.} \quad x^m, d^{n,m} \geq 0 \tag{3.7}$$

where β is a scalar which controls how sparse the resulting warping matrix is. In this cost function, there is a second regularization term which penalizes the l_1-norm of the rows of the warping matrix to induce sparsity. It should be noted that the structural sparsity of the warping matrix limits the freedom in time warping by allowing only a few consecutive frames with nonzero weights, whereas the regularized sparsity implies that the linear approximation is dominated by a single time-frequency cell obtaining a much larger weight compared to the others. To minimize the cost function in Equation (3.7), the multiplicative update rules given below are applied iteratively,

$$\mathbf{x} \leftarrow \mathbf{x} \odot ((\mathbf{D}\mathbf{S})^T (\mathbf{y}_{l_i} \oslash \mathbf{D}\mathbf{S}\mathbf{x})) \oslash ((\mathbf{D}\mathbf{S})^T \mathbf{1}_x + \Lambda) \tag{3.8}$$

$$\mathbf{D} \leftarrow \mathbf{D} \odot ((\mathbf{y}_{l_i} \oslash \mathbf{D}\mathbf{S}\mathbf{x})(\mathbf{S}\mathbf{x})^T) \oslash (\mathbf{1}_D (\mathbf{S}\mathbf{x})^T + \beta) \tag{3.9}$$

with $\odot$ and $\oslash$ denoting element-wise multiplication and division respectively. $\mathbf{1}_x$ is a Fl_e-dimensional vector and $\mathbf{1}_D$ is a Fl_i-dimensional vector with all elements equal to unity. After each iteration, the rows of the warping matrix $\mathbf{D}$ are normalized to unity in order to avoid extremely small or large values in $\mathbf{D}$ and $\mathbf{x}$. Applying these update rules iteratively, $\mathbf{D}$ and $\mathbf{x}$ become sparser and the reconstruction error between the input speech vector and its approximation decreases monotonically. A reconstruction error-based decoding is applied to find the best matching class sequence using dynamic programming. A known problem of sparse representation approaches working on magnitude spectra is

that the silence exemplars are not recognized [50]. This is due to the fact that silence is well-approximated by combining speech exemplars with small weights, so all classes will score equally well. To overcome this problem, reconstruction errors for the class representing silence have to be compensated. The details of the reconstruction error-based decoding and silence dictionary scoring can be found in [187].

3.2.3 Designing the Warping Matrix

A warping function is defined as a mapping between the time axes of two different patterns (exemplars and input speech segments in this case) [145]. Such a function is expected to capture the spectral similarities between two frame sequences with different durations. To prevent unnatural mappings, some conditions are imposed on the warping function. The warping matrix discussed in Section 3.2.2 should be properly designed so that it also satisfies these warping function conditions, namely monotonicity, continuity, boundary, adjustment window and slope constraint conditions, which are defined in [145]. Monotonicity and continuity conditions prohibit warping backwards and limit the number of skipped or stalled frames for two consecutive input speech frames. Boundary condition implies matching the first and last frame with the first and last input speech frame respectively. The adjustment window constraint and slope constraint conditions aim to confine the warping path by preventing too many successive skips or stalls.

A warping matrix $\mathbf{D}$ of dimensionality $Fl_i \times Fl_e$ linearly combines the corresponding time-frequency cells belonging to consecutive frames in $\hat{\mathbf{y}}_{l_e}$ to approximate Fl_i input time-frequency cells. Considering the aforementioned conditions, the initial $\mathbf{D}$ matrix is composed of identity submatrices $\mathbf{I}$ of dimensionality $F \times F$ on the diagonal and either sub- or superdiagonal depending on the sign of $l_i - l_e$. For the case of $l_i = l_e + 1$,

$$\mathbf{D} = \begin{bmatrix} \mathbf{I} & 0 & 0 & \cdots & 0 & 0 \\ \mathbf{I} & \mathbf{I} & 0 & \cdots & 0 & 0 \\ 0 & \mathbf{I} & \mathbf{I} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{I} & 0 \\ 0 & 0 & 0 & \cdots & \mathbf{I} & \mathbf{I} \\ 0 & 0 & 0 & \cdots & 0 & \mathbf{I} \end{bmatrix} \tag{3.10}$$

and $l_i = l_e - 1$,

$$\mathbf{D} = \begin{bmatrix} \mathbf{I} & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \mathbf{I} & \mathbf{I} & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \mathbf{I} & \mathbf{I} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \mathbf{I} & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & \mathbf{I} & \mathbf{I} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \mathbf{I} \end{bmatrix} \tag{3.11}$$

The design can be generalized to any l_i and l_e, once the warping matrix satisfies the warping conditions.

3.3 Experimental Setup

3.3.1 Database

The exemplars used in experiments are speech segments extracted from the clean training set of AURORA-2 database [77] which contains 8440 utterances with one to seven digits in American English. There are 4 clean test sets, each containing 1001 utterances and recognition experiments are performed on these test sets.

3.3.2 Baseline System

Exemplars and input speech segments are represented in root-compressed (with magnitude power = 0.66) mel-scaled magnitude spectra. A 17 channel mel-scaled filter bank with triangular magnitude response is computed from a spectral analysis with a frame length of 32 ms and a frame shift of 10 ms. The first channel is centered at 200 Hz and the last is at 3030 Hz.

The training data is segmented into the exemplars representing half-digits by a conventional HMM-based recognizer. The system uses 508 dictionaries belonging to 23 different classes. The minimum and maximum exemplar lengths are 5 and 30 frames respectively. Exemplars longer than 30 frames are removed to limit the number of dictionaries. The baseline system uses 10,362 exemplars in total including 260 silence exemplars. Λ is set to 2. The l_2-norm of each dictionary column and reshaped input speech vectors are normalized to unity. The reconstruction error shows enough discrimination among different classes after 50 iterations. Further details about the baseline system can be found in [187].

Table 3.1: Average word error rates obtained on four clean test sets (SR: Sparse representations, TW: Time warping)

	WER (%)
SR (baseline)	1.91
SR + TW	1.78
SR + TW + Sparsity ($\beta = 10$)	1.66
SR + TW + Sparsity ($\beta = 20$)	1.64
SR + TW + Sparsity ($\beta = 100$)	1.66

3.3.3 Implementation Details

The proposed system is implemented in MATLAB and GPUs are used to accelerate the evaluation of Equation (3.8) and (3.9). We have not made the effort yet to design a dedicated implementation exploiting the sparse structure of the warping matrix $\mathbf{D}$, i.e. in our current implementation the zero entries in $\mathbf{D}$ are reestimated as well. Avoiding this is expected to reduce the simulation times significantly, but requires a significant software engineering effort on a GPU, which has not been performed to date.

3.4 Results and Discussion

This section presents the preliminary recognition results obtained using the proposed sparse representation model with time warping. The experiments put more focus on the impact of sparsity regularization imposed on the warping matrix rather than the relative performance of different warping matrix designs. The recognition is performed by approximating input speech segments of length l_i by linearly combining the exemplars of length $l_e = l_i, l_i \pm 1$ using the warping matrices discussed in Section 3.2.3. These warping matrices linearly combine time-frequency cells belonging to two successive frames to approximate input speech frames except for the first and last input speech frames.

The baseline system uses the sparse representation model described in [187]. The WER obtained with the baseline system is 1.91% which is given in the first row of Table 3.1. The average simulation time for the baseline system is approximately 3 seconds/utterance. The proposed model with $\beta = 0$ performs better than the baseline with a WER of 1.78% given in the second row. This improvement comes with a great increase in the average simulation time mostly due to the higher number of matrix multiplications in the multiplicative update rules given in Equation (3.8) and (3.9). Recognition of each utterance using the proposed

model takes 45 seconds on average. After setting β to several nonzero values, $\beta = 10, 20$ and 100 in this case, the WER further reduces to 1.64% for $\beta = 20$. This result shows the positive impact of imposing sparsity regularization on the warping matrix combined with the structural sparsity. This is due to the fact that one of the two time-frequency cells in the consecutive frames gets a much higher weight than the other resulting in a realistic approximation of the input time-frequency cell. Furthermore, it is evident that the recognition accuracy does not vary significantly for different β values. The results discussed above prove the feasibility of the proposed model providing 14% relative improvement in the WER with time warping limited to a single frame.

The time warping technique we have proposed is different from classical DTW in several aspects. The main difference is that the proposed time warping scheme learns distinct weights for each time-frequency cell whereas classical DTW provides a frame-level mapping between the time axes. One way of adopting a frame-level mapping in the proposed framework is to tie the time-frequency cell weights which belong to the same frame, a constraint for which new multiplicative update formulae have been derived and which will be evaluated in our future work.

Another difference is that classical DTW applies dynamic programming to obtain a warping path through the time axes of the different-length segments. In our case, the complete warping path is learned by fitting a product of matrices to the data. Finally, the conditions on the warping function are imposed more explicitly in classical DTW compared to the proposed approach. The only way to impose these conditions in the proposed scheme is the careful design of the warping matrix. Even with a carefully designed warping matrix, it is not possible to implement some slope constraints such as the Itakura constraint [88].

3.5 Conclusions and Future Work

In this chapter, we have introduced a novel sparse representation model for speech signals which allows time warping. This model approximates input speech segments as a product of three matrices, i.e. a sparsely structured warping matrix that linearly combines the time-frequency cells of consecutive frames, a dictionary containing exemplars that are extracted from training data and a weight vector storing the exemplar weights. The design of the warping matrix is of great importance to obtain realistic warping paths. Two warping matrices are introduced for matching two frame sequences with a single frame difference.

Applying this model to recognize digit sequences, we analyze the impact of inducing sparsity in the warping matrix by penalizing the l_1-norm of the rows of the warping matrix. The results have shown that the proposed sparse representation model allowing time warping provides 7% relative improvement in the WER compared to a baseline system which compares input speech segments and exemplars of the same length only. Moreover, the existence of sparsity regularization improves the recognition further yielding a total relative improvement of 14%. These improvements come with a cost of higher computational complexity increasing the average recognition time by a factor of 15, though this number should be interpreted with care given the current sub-optimal implementation.

Chapter 4

Speech Exemplar Selection Techniques from Multiple Dictionaries

This chapter describes and analyzes several exemplar selection techniques to reduce the number of exemplars that are used in a recently proposed sparse representations-based speech recognition system. Exemplars are labeled acoustic realizations of different durations which are extracted from the training data. For practical reasons, they are organized in multiple undercomplete dictionaries, each containing exemplars of a certain speech unit. Using these dictionaries, the input speech segments are modeled as a sparse linear combination of exemplars. The improved recognition accuracy with respect to a system using fixed-length exemplars in a single dictionary comes with a heavy computational burden. Due to this fact, we investigate the performance of various exemplar selection techniques that reduce the number of exemplars according to different criteria and discuss the links between the salience of the exemplars and the data geometry. The pruned dictionaries using only 30% of the exemplars have been shown to achieve comparable recognition accuracies to what can be obtained with the complete dictionaries.

This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Exemplar Selection Techniques for Sparse Representations of Speech Using Multiple Dictionaries*", In 21st European Signal Processing Conference (EUSIPCO), pages 1-5, Marrakesh, Morocco, Sept. 2013.

4.1 Introduction

The success of recently proposed speech recognition systems based on template matching attracted considerable interest in exemplar-based acoustic modeling as a viable alternative to its statistical counterparts [30, 143]. Exemplars are labeled speech segments such as phones, syllables or words, possibly of different length, that are extracted from the training data. Each exemplar is tagged with meta-information including speaker and environmental characteristics. An input speech segment can simply be classified by evaluating the labels of the spatially closest exemplars. Inconsistent exemplar sequences, e.g. sequences with different gender exemplars, can be penalized based on the tagged meta-information.

Although exemplars provide better duration and trajectory modeling compared to the Hidden Markov Models, large amounts of data are required to handle the acoustic variation among different utterances [30]. In order to reduce high memory and computational power requirements, several exemplar selection algorithms are proposed in [149, 158]. The main goal of these techniques is to remove less informative exemplars that are hardly used or whose presence result in inaccurate recognition and to achieve comparable recognition accuracies using only a portion of the exemplars.

Another framework in exemplar-based techniques, namely exemplar-based sparse representations (SR), models the spectrogram of input speech segments as a sparse linear combination of exemplars rather than comparing with each individual exemplar. SR-based techniques have been successfully used for speech enhancement [54], feature extraction [142] and clean [48] and noisy [49, 84, 162] speech recognition. In these approaches, fixed-size exemplars are stored in the columns of an overcomplete dictionary which has much higher number of columns (exemplars) than rows (time-frequency cells). We have recently proposed an SR-based speech recognition system which uses exemplars of different length organized in separate dictionaries and which approximates the input speech as a linear combination of the exemplars in each dictionary [187]. Most of these dictionaries are undercomplete having less exemplars than the number of time-frequency cells. We have also shown that this system performs reasonably well under noisy conditions in [181].

Reducing the dimensions of large datasets stored in overcomplete dictionaries has been investigated in different fields and several matrix decompositions such as the singular value decomposition (SVD), rank revealing QR decomposition, CUR matrix decomposition, interpolative decomposition (ID) have been used to obtain a low-rank matrix approximation of the complete data matrix [64]. Though the SVD is known to provide the best rank-k approximation, interpretation of the principal components is difficult in data analysis [115]. Therefore, several CUR

matrix decompositions have been proposed in which a matrix is decomposed as a product of three matrices $\mathbf{C}$, $\mathbf{U}$, $\mathbf{R}$ and the matrices $\mathbf{C}$ and $\mathbf{R}$ consist of a subset of the actual columns and rows respectively [38, 44, 61]. Moreover, a probabilistic ID technique which automatically handles the model selection is introduced and applied to a polyphonic music transcription task using an overcomplete dictionary containing exemplars of different musical notes in [4].

The exemplar selection techniques proposed in this chapter differ from previous work as the dictionaries, which only contain exemplars of the same length and label, are undercomplete due to insufficient training data. Compared to the overcomplete dictionaries with a large number of data points, the redundancy in undercomplete dictionaries is quite limited. Therefore, removing a few highly relevant data points may already result in significant decreases in the recognition accuracy. The use of real exemplars tagged with meta-information is another requirement which prevents applying the SVD or any clustering technique. To the best of our knowledge, there is no prior work on selecting the most salient columns of an undercomplete dictionary. In this chapter, we propose various techniques for selecting the most informative columns of the undercomplete dictionaries and analyze the selection problem elaborating on the geometrical structure of the data.

The rest of the chapter is organized as follows. A brief description of the sparse representations-based speech recognition system is given in Section 2. The proposed exemplar selection techniques are discussed in Section 3. Section 4 explains the experimental setup and implementation details. In Section 5, we present the recognition results and a general discussion on the proposed techniques is given. The conclusions and thoughts for future work are discussed in Section 6.

4.2 System Description

The recognition system that is described in [187] uses a sparse linear combination of the exemplars to model the input speech segments. Each exemplar is associated with a certain speech unit and the duration of each speech unit in the training data is preserved yielding exemplars of different lengths.

Exemplars spanning l frames are reshaped into a single vector and stored in the columns of the dictionary $\mathbf{S}_{c,l}$: one for each speech unit c and each length l. Each dictionary is of dimensionality $Dl \times N_{c,l}$ where D is the number of frequency bands in a frame and $N_{c,l}$ is the number of available exemplars of length l and class c. For any class c, a reshaped input speech vector $\mathbf{y}_l$ of length Dl is expressed as a linear combination of the exemplars with non-negative

weights:

$$\mathbf{y}_l \approx \sum_{m=1}^{N_{c,l}} x_{c,l}^m \mathbf{s}_{c,l}^m = \mathbf{S}_{c,l}\mathbf{x}_{c,l} \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \tag{4.1}$$

where $\mathbf{x_{c,l}}$ is an $N_{c,l}$-dimensional sparse weight vector. Sparsity of the weight matrix implies that the input speech is approximated by a small number of exemplars. The exemplar weights are obtained by minimizing the cost function,

$$d(\mathbf{y}_l, \mathbf{S}_{c,l}\mathbf{x}_{c,l}) + \Lambda \sum_{m=1}^{N_{c,l}} x_{c,l}^m \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \tag{4.2}$$

where Λ is a scalar which controls how sparse the resulting vector $\mathbf{x}$ is. The first term is the divergence measure between the input speech vector and its approximation. The second term is a regularization term which penalizes the l_1-norm of the weight vector to produce a sparse solution. The generalized Kullback-Leibler divergence (KLD) is used for d:

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k \tag{4.3}$$

The regularized convex optimization problem can be solved using various methods including non-negative sparse coding (NSC). For NSC, the multiplicative update rule to minimize the cost function (4.2) is derived in [50] and is given by

$$\mathbf{x}_{c,l} \leftarrow \mathbf{x}_{c,l} \odot (\mathbf{S}_{c,l}^T(\mathbf{y}_l \oslash (\mathbf{S}_{c,l}\mathbf{x}_{c,l}))) \oslash (\mathbf{S}_{c,l}^T\mathbf{1} + \Lambda) \tag{4.4}$$

with $\odot$ and $\oslash$ denoting element-wise multiplication and division respectively. $\mathbf{1}$ is a Dl-dimensional vector with all elements equal to unity. Applying this update rule iteratively, the weight vector becomes sparser and the reconstruction error between the input speech vector and its approximation decreases monotonically.

The first term of Equation (4.2) expresses the reconstruction error between a speech segment of length l and a class c. Every speech segment of each available exemplar length is approximated as a linear combination of exemplars. This is achieved by applying the sliding window [50] to the input utterance for each available exemplar length and iteratively applying equation (4.4) using the dictionaries of the corresponding length. After a fixed number of iterations, the reconstruction error is calculated. As the label of each dictionary is known, decoding is performed by finding the class sequence that minimizes the reconstruction error using dynamic programming.

4.3 Exemplar Selection Techniques

The computational bottleneck of the system described above is the evaluation of Equation (4.4). The computational complexity per iteration is linearly proportional to the number of exemplars and it can be reduced by removing the less informative and redundant exemplars that are either not used or result in misclassifications. The baseline column selection technique is the randomized column selection algorithm which is proposed as a part of the CUR matrix decomposition in [115]. This algorithm randomly selects a subset of the columns of a data matrix with respect to the probability distribution computed as the normalized statistical leverage scores. Preferably selecting high-statistical leverage columns will, with high probability, lead to a reduced dictionary which approximates the original one almost as well as an SVD-based rank reduction scheme [115].

In this section, we propose several exemplar selection techniques that reduce the number of exemplars stored in the dictionaries discussed in Section 4.2. These techniques are classified into three categories, namely reconstruction error-based, distance-based and activation-based according to their exemplar selection criteria.

4.3.1 Reconstruction error-based techniques

The system described in Section 4.2 approximates input segments as a linear combination of exemplars. Since the approximation quality is measured using the divergence measure in Equation (4.3), the approximation of an exemplar either using other exemplars in the same-class dictionary or the ones in different-class dictionaries of the same length provides useful information about its salience.

Collinearity reduction (CR)

Exemplars that are well approximated by the other exemplars from the same-class dictionary contain less information compared to the ones with higher reconstruction errors. Therefore, the collinearity reduction technique removes the exemplars that are well approximated as a linear combination of the other exemplars in the same-class dictionary with non-negative weights. This idea is applied iteratively by removing the exemplar that is approximated with the minimum reconstruction error at each iteration until the minimum number of exemplars requirement in a dictionary is met.

Discriminative dictionaries (DD)

Dictionary elements of a particular class that are well approximated by a dictionary of another class are likely to cause confusion during recognition. Indeed, any data that is close to these elements may be explained as belonging to the wrong class. The discriminative dictionaries technique iteratively removes the exemplars having the smallest ratio between the reconstruction errors that are obtained using the dictionary containing the exemplars of the other classes and the same-class dictionary.

4.3.2 Distance-based techniques

Distance-based techniques perform exemplar selection considering the spatial closeness of the exemplars which provides information about the data geometry. The symmetric KLD is used as a distance metric which is defined as

$$d_{skld}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2}(d(\mathbf{y}, \hat{\mathbf{y}}) + d(\hat{\mathbf{y}}, \mathbf{y})) \tag{4.5}$$

where d is defined in Equation (4.3).

Removing exemplars with the smallest/largest average distance (SAD/LAD)

Removing the same-class exemplars that either lie in the densely or sparsely populated regions in the feature space has been investigated. This technique retains the exemplars having either the smallest or the largest average distance to the other exemplars stored in the same-class dictionary.

Pruning the closest exemplars (CE)

The second distance-based technique aims to reduce the number of exemplars by discarding one of the exemplars that lie close to each other. At each iteration, the two closest exemplars are found and only one of them is retained in the dictionary.

4.3.3 Activation-based technique

Active exemplars (AE)

A single activation-based technique is proposed which infers the salience of an exemplar by evaluating the average weight it gets on a recognition task. The exemplar weights in the described system are obtained by applying the multiplicative update rule in Equation (4.4). Obviously, the exemplars often having higher weights are more decisive in the recognition. Thus, less *active* exemplars are rarely used and they can be removed from the dictionary without a significant loss in the recognition accuracy. The training data is used to quantify how active each exemplar is.

4.4 Experimental Setup

4.4.1 Database

The exemplars used in experiments are speech segments extracted from the clean training set of AURORA-2 database [77] which contains 8440 utterances with one to seven digits in American English. The performance of the proposed exemplar selection techniques is evaluated on the clean test sets of the same database. There are 4 clean test sets, each containing 1001 utterances and recognition experiments are performed on these test sets using the pruned dictionaries.

4.4.2 Baseline System

Exemplars and input speech segments are represented in root-compressed (with magnitude power = 0.66) mel-scaled magnitude spectra with 17 frequency bands. The frame length is 32 ms and the frame shift is 10 ms. The training data is segmented into the exemplars representing half-digits by a conventional HMM-based recognizer. The system uses 508 dictionaries belonging to 23 different classes. The largest number of exemplars in a dictionary is 283. The minimum and maximum exemplar lengths are 5 and 30 frames respectively. Exemplars longer than 30 frames are removed to limit the number of dictionaries. The baseline system uses 50,654 exemplars in total including 1300 silence exemplars. The l_2-norm of each dictionary column and reshaped input speech vectors are normalized to unity. Further details about the baseline system can be found in [187].

Table 4.1: Average word error rates obtained on four clean test sets using the complete and pruned dictionaries. The first row provides the result obtained using the complete dictionaries.

Removed exemplars (%)	# of exemplars	CR	AE	SAD	CE	DD	LAD	CUR
0	50654	1.68	1.68	1.68	1.68	1.68	1.68	1.68
10	45968	1.66	1.67	1.67	1.72	2.05	2.30	1.67
20	40858	1.73	1.69	1.69	1.76	4.43	2.71	1.57
30	35793	1.79	1.76	1.69	1.78	2.73	3.10	1.51
40	30687	1.75	1.73	1.69	1.81	2.99	3.62	**1.73**
50	25531	1.76	1.75	1.69	**1.78**	3.41	4.15	1.97
60	20533	1.76	1.77	**1.79**	1.92	3.86	4.51	2.01
70	15468	**1.79**	**1.84**	2.08	2.01	4.29	4.90	2.10
80	10362	1.91	2.30	2.19	2.14	5.27	5.92	2.50
90	5293	2.28	4.66	3.80	2.58	6.87	6.77	3.18

4.4.3 Implementation of the Proposed Techniques

All of the proposed techniques are applied before the recognition experiments to create the pruned dictionaries. Reconstruction error and activation-based techniques require the evaluation of the multiplicative update rule given in Equation (4.4) in order to obtain the exemplar weights. The CR and DD techniques are applied iteratively discarding a single exemplar at each step. The AE technique, on the other hand, stores the average weight each exemplar gets during the approximation of the speech segments from the training data and the exemplar selection is performed by preserving the required number of exemplars with the highest average weight value. Distance-based techniques use a square and symmetric distance matrix to identify the spatial closeness of the exemplars. The CE technique iteratively reduces the number of exemplars while the DP and SP techniques are applied in a single step. The recognition accuracies presented in the following section are obtained by reducing the number of exemplars in each dictionary by 10% at each step until only 10% of the exemplars remain in each dictionary.

4.5 Results and Discussion

In this section, we present the word error rates (WER) that are obtained on the clean test set of AURORA-2 using the dictionaries pruned with the techniques discussed in Section 4.3. These results are compared with the recognition accuracies obtained with the complete dictionaries and the dictionaries that are

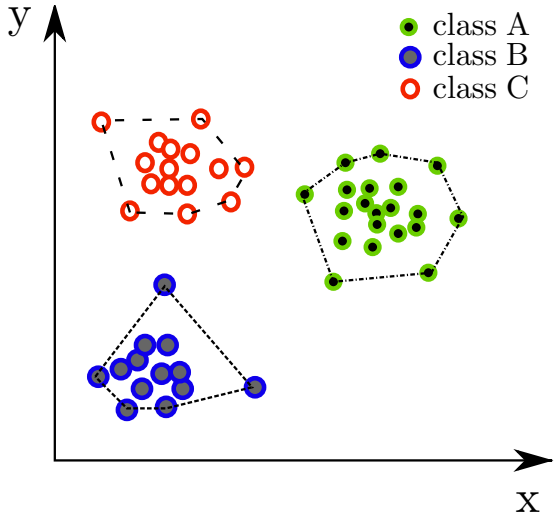


Figure 4.1: Illustration of the convex hulls formed by the same class exemplars in two dimensions.

pruned with the randomized column selection algorithm of the CUR matrix decomposition. The recognition experiments on clean data provide information about both the performance of the proposed exemplar selection techniques and the size of the smallest dictionaries that sufficiently represent clean speech as a design parameter. It is worth mentioning that basic HMM/GMM systems provide higher recognition accuracies (about a percent) on the clean test set compared to the baseline recognizer using the complete dictionaries [77]. However, unlike this framework, it is not easy to account for background noise in HMMs [181].

Table 4.1 presents the WER results. The baseline system with the complete dictionaries has a WER of 1.68%. For each technique, the smallest dictionary size in which the WER has increased less than 10% (i.e. 1.68% * 1.1 = 1.85%) over the baseline is given in bold. The dictionaries pruned with collinearity reduction (CR) and active exemplars (AE) provide results lying in this error bound using 30% of the exemplars. Removing the exemplars with the smallest average distance (SAD) and pruning the closest exemplars (CE) performs slightly worse than the CR and AE staying in the bound using 40% and 50% of the exemplars respectively. The CUR decomposition gives similar WERs using more than 50% of the exemplars. The simulation times of the final system using 30% of the exemplars are reduced by a factor of 3, varying from 2.8 to 4 seconds depending on the utterance duration.

Table 4.2: Average word error rates obtained on four clean test sets using the DD and LAD techniques for outlier removal.

Removed exemplars (%)	# of exemplars	DD	LAD
0	50654	1.68	1.68
1	50568	1.76	1.72
2	50045	1.83	1.82
3	49544	1.85	1.95
4	49018	1.85	2.06

The hypothetically appealing idea of obtaining more discriminative dictionaries (DD) and removing the exemplars with the largest average distance (LAD) do not work for the intended task. Even after removing 10% of the exemplars, the WER exceeds 2%. The results obtained with these techniques imply that the spatial position of a data point provides some clues about how informative it is in the recognition. Due to the non-negativity of the data, each dictionary forms a convex hull that lies in the positive orthant. There are a few exemplars that lie on or next to the boundaries whereas the center is densely populated. A two dimensional illustration of the ideal (perfectly separable) case with three different classes is given in Figure 4.1. Considering the exemplar selection criteria of the LAD, it is apparent that it mainly discards exemplars that are further away from the densely populated region in the convex hulls. Similarly, the DD aims to reduce the confusions between the dictionaries and these confusions are mostly due to the exemplar lying on the boundaries in each convex hull. Removing these exemplars results in narrower convex hulls spanned by each dictionary which provides a less accurate description of the cone. On the other hand, other techniques retaining the exemplars lying on the boundaries and preserving the convex hull formed by each dictionary performs significantly better than the DD and LAD. It should be noted that most active exemplars typically lie on the convex hull boundaries which are rather decisive in the recognition.

Although the importance of the exemplars lying on the boundaries for the recognition accuracy has been shown, it can still be claimed that some of these exemplars can be outliers resulting in misclassifications. A discussion on the misclassifications due to the outliers in a convex hull can be found in [141]. To analyze the impact of the outliers on the recognition accuracy, we further apply the DD and LAD to remove a few percent of the exemplars. From the results in Table 4.2, it is not evident that these techniques work for outlier removal either. This can be either due to the non-existence of outliers in most dictionaries or their negligible impact on the recognition accuracy.

4.6 Conclusions and Future Work

In this chapter, we have proposed several exemplar selection techniques for undercomplete dictionaries and analyzed which exemplars these techniques tend to select considering the geometrical structure formed by the data points in the feature space. Techniques based on the collinearity reduction (CR) and selecting the active exemplars (AE) provided the best results by achieving recognition accuracies that are in the 10% error bound of the baseline results using only 30% of the exemplars. The distance-based techniques, namely removing exemplars with the smallest average distance (SAD) and pruning the closest exemplars (CE), perform slightly worse than the CR and AE. All of these techniques outperform the CUR decomposition which has been successfully used for reducing the size of overcomplete dictionaries.

Discriminative dictionaries (DD) and removing the exemplars with the largest average distance (LAD) provide inferior results revealing the connection between the spatial position of an exemplar and its salience in the recognition. The DD and LAD mostly discard exemplars lying on the boundaries of the convex hulls resulting in a less accurate description of the cone. On the other hand, the SAD and CE explicitly remove the exemplars lying in the densely populated region of the convex hulls without deforming the boundaries and provide much better results than the DD and LAD. Hence, it can be concluded that the exemplars lying on the boundaries of the convex hulls are highly informative and discarding these exemplars results in high recognition accuracy loss.

Chapter 5

Noise Robust Exemplar Matching (N-REM) for ASR

Performing automatic speech recognition using exemplars (templates) holds the promise to provide a better duration and coarticulation modeling compared to conventional approaches such as hidden Markov models (HMM). Exemplars are spectrographic representations of speech segments extracted from the training data, each associated with a speech unit, e.g. phones, syllables, half-words or words, and preserve the complete spectro-temporal content of the speech. Conventional exemplar-matching approaches to automatic speech recognition systems, such as those based on dynamic time warping, have typically focused on evaluation in clean conditions. In this chapter, we propose a novel noise robust exemplar matching framework for automatic speech recognition. This recognizer approximates noisy speech segments as a weighted sum of speech and noise exemplars and performs recognition by comparing the reconstruction errors of different classes with respect to a divergence measure. We evaluate the system performance in keyword recognition on the small vocabulary track of the 2^{nd} CHiME Challenge and connected digit recognition on the AURORA-2 database. The results show that the proposed system achieves comparable results with state-of-the-art noise robust recognition systems.

This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Noise Robust Exemplar Matching Using Sparse Representations of Speech*", IEEE/ACM Transactions on Audio, Speech, and Language Processing, volume 22, No. 8, pages 1306-1319, Aug. 2014.

5.1 Introduction

Using segments of speech that are extracted from training data (exemplars) for automatic speech recognition (ASR) has seen renewed interest on account of the huge increase in computational power and fast template matching algorithms [30, 34, 143]. Rather than learning acoustic models which are based on distributions (e.g. hidden Markov models (HMM)) or latent variables (e.g. deep neural networks), exemplar-based acoustic modeling uses the data itself to explain unseen speech. Well-known issues of the former parametric acoustic models such as inaccurate duration modeling, limited coarticulation modeling and overgeneralization may be circumvented in exemplar-based acoustic modeling since the long-range spectro-temporal dynamics of speech can be preserved.

Prior work on ASR systems employing exemplar matching, i.e. classifying an observed speech segment using the label(s) of the closest exemplar(s), investigates the recognition performance under clean conditions. In this chapter we focus on the noise robustness of the exemplar matching approach, and propose a novel exemplar-matching framework, dubbed *noise robust exemplar matching* (N-REM), that allows noise modeling. The proposed approach does not rely on preprocessing (e.g. feature/model compensation) or on postprocessing (e.g. uncertainty decoding). Instead, recognition works by approximating the spectral representations of noisy speech segments as a superposition of speech and noise exemplars and performing reconstruction error-based decoding by applying dynamic programming.

Exemplars are labeled speech segments associated with a single speech unit such as phones, syllables, half-words or words that have occurred in the training data. Every speech unit has a distinct duration distribution resulting in exemplars spanning multiple lengths. They are typically compared with the observed speech segments using dynamic time warping (DTW) [29,135,144], and an input speech segment can be classified as the label of either the closest exemplar or by performing a majority voting on the set of K nearest neighbors [36, 57, 160]. Several hybrid recognition systems which combine this approach with statistical models have also been proposed [1, 9, 28, 55, 157].

Applying exemplar matching under noisy conditions creates training-test set mismatch problems similar to what is experienced with conventional ASR systems [176]. As for those systems, feature compensation methods can be applied to increase the noise robustness without modifying the recognizer. Model compensation techniques would require all exemplars to be modified during decoding, which is a formidable task in the case of non-stationary noise. Since the search problem in exemplar-based recognition is a lot more involved than in conventional ASR systems based on HMMs, the equivalent of factorial

models does also seem infeasible. Finally, multi-condition training, i.e. storing noisy exemplars, will increase the number of exemplars that need to be stored dramatically. Furthermore, noisy exemplars would only capture a single instance of speech and noise resulting in a limited noise modeling capacity especially in case of non-stationary noise.

Recently, speech processing based on *sparse representations* (SR), has been successfully used for speech enhancement [54], feature extraction [142] and clean [48] and noisy [49, 84, 162] speech recognition. In this approach, observed speech spectra are modeled as a sparse linear combination of atoms describing parts of spectra, organized in an overcomplete dictionary. By containing speech and noise atoms in a single dictionary, noise robustness is achieved by modeling noisy speech explicitly as a superposition of both speech and noise atoms. In [48, 49, 54, 84], these speech and noise atoms consisted of exemplars, as it allowed for accurate modeling of temporal context. Unlike their use in exemplar-matching ASR systems, these fixed-length exemplars are randomly extracted and do not model a specific choice of speech units.

The main contribution of this work is a new exemplar matching recognition framework that allows noise modeling. Based on preliminary research reported in [181, 183, 187], N-REM uses an SR-based exemplar matching approach with exemplars of multiple length corresponding to speech units. The exemplars are organized in separate dictionaries based on the length and class (associated speech unit), and are used to approximate noisy speech segments as a linear combination of the exemplars in each of these dictionaries. Non-negative sparse coding (NSC) is applied to determine the weights of the linear combination [49]. The recognizer adopts a reconstruction error-based back-end, i.e. the recognition is performed by comparing the quality of the match for different classes quantified by a distance/divergence measure and choosing the class sequence that minimizes the total reconstruction error.

The development of the novel framework also involves a dedicated design of the dictionaries that takes computational limitations into account. In previous work, we have proposed several dictionary design techniques for speech dictionaries, focused at either reducing the dictionary sizes for overpopulated dictionaries [182], or at increasing the number of exemplars available to improve the acoustic modeling of underpopulated dictionaries [183]. In this work, we propose a novel adaptive noise dictionary design technique. This technique adaptively selects a small number of noise exemplars that are expected to model the actual noise conditions.

The rest of the chapter is organized as follows. The proposed exemplar-matching framework scheme is described in Section 5.2. The design techniques that are applied to the dictionaries are described in Section 5.3. The evaluation setup is

described in Section 5.4. The recognition results on the small vocabulary track of the 2nd CHiME Challenge and the AURORA-2 database are presented and discussed in Section 5.5. A general discussion on the recognition performance and some future directions are given in Section 5.6. Section 5.7 concludes the chapter.

5.2 Sparse Representation Model of Speech with Exemplars of Multiple Length

5.2.1 Modeling noisy speech

N-REM models noisy speech segments as a sparse linear combination of speech and noise exemplars of various lengths that are stored in multiple dictionaries. The overview of the recognizer is given in Figure 5.1. Compared to a system using fixed-length exemplars stored in a single dictionary, using separate dictionaries for each class results in noisy speech segments being approximated as a linear combination of exemplars belonging to the same class only. From the geometrical interpretation of NSC-based source separation, it is known that the farther the convex hull of the basis vectors belonging to different sources (speech and noise in this case) are, the better the separation is [37]. Hence, the use of separate dictionaries for each speech unit provides a more precise representation in the high-dimensional feature space.

Training frame sequences representing various speech units (speech exemplars) are extracted based on the state-level alignments obtained using a conventional HMM-based recognizer. Speech exemplars, each comprised of D mel frequency bands and spanning l frames, are reshaped into a single vector and stored in the columns of a speech dictionary $\mathbf{S}_{c,l}$: one for each class c and each length l. Each dictionary is of dimensionality $Dl \times N_{c,l}$ where $N_{c,l}$ is the number of available speech exemplars of class c and length l. Similarly, a single noise dictionary $\mathbf{N}_l$ for each length l is formed by reshaping the noise exemplars. Each speech dictionary is concatenated with the noise dictionary of the same length to form a combined dictionary $\mathbf{A}_{c,l} = [\mathbf{S}_{c,l}\, \mathbf{N}_l]$ of dimensionality $Dl \times M_{c,l}$ where $M_{c,l}$ is the total number of available speech and noise exemplars.

An observed noisy (and/or reverberated) speech segment of length T frames is also reshaped into vectors by applying a sliding window approach [50] with window length of l frames and stored in an observation matrix $\mathbf{Y}_l = [\mathbf{y}_l^1, \mathbf{y}_l^2 ..., \mathbf{y}_l^{(T-l+1)}]$ of dimensionality $Dl \times (T - l + 1)$. Due to multiple-length exemplars, the window length l is varied between the minimum exemplar length $l_{\min}$ and maximum exemplar length $l_{\max}$ yielding observation matrices $\mathbf{Y}_l$ for

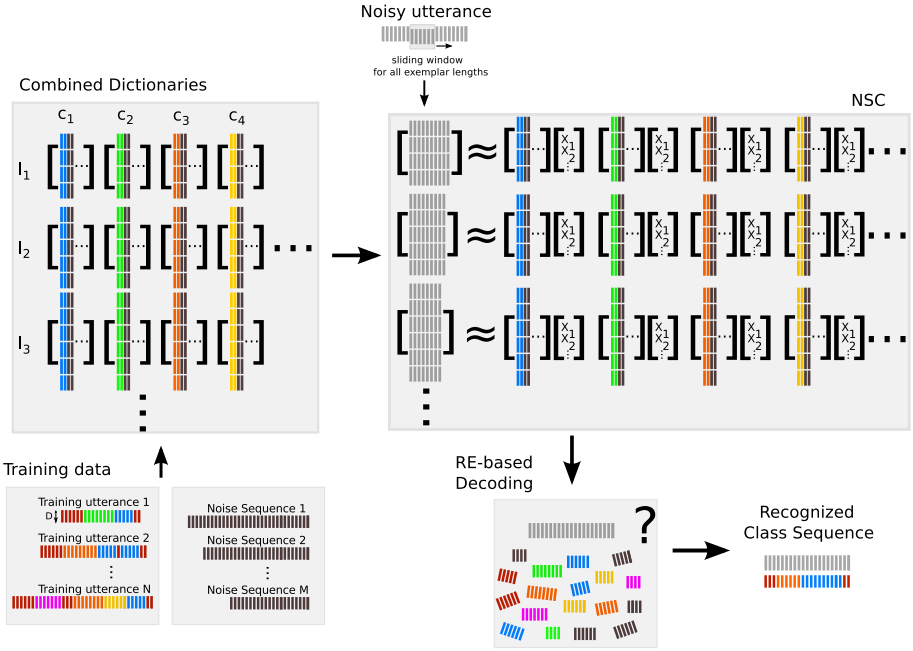


Figure 5.1: The Recognizer Overview. Speech exemplars are extracted from the training data using the segmentation information. They are organized in dictionaries based on their length and class (associated speech unit). Noise dictionaries are concatenated to the speech dictionaries forming the combined dictionaries. Non-negative sparse coding (NSC) is applied to approximate noisy test utterances using the combined dictionaries. After a fixed number of iterations, the reconstruction errors are calculated and a single-stage dynamic programming algorithm is applied to find the class sequence with the minimum reconstruction error as the dictionary labels are known.

$l_{\min} \leq l \leq l_{\max}$. For every class c, each observation vector $\mathbf{y}_l$ is expressed as a linear combination of the exemplars that are stored in the dictionaries of the same length:

$$\mathbf{y}_l \approx \sum_{m=1}^{M_{c,l}} x_{c,l}^m \mathbf{a}_{c,l}^m = \mathbf{A}_{c,l}\mathbf{x}_{c,l} \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \tag{5.1}$$

where $\mathbf{x}_{c,l}$ is an $M_{c,l}$-dimensional non-negative weight vector. The sparse solutions of $\mathbf{x}_{c,l}$ yield more realistic approximations of the observed segments without overfitting and have been shown to provide better recognition results [80, 174].

The combined dictionaries consisting of speech and noise exemplars are presumed to model all acoustic variability in the observed signal due to pronunciation variation, background noise and so forth. This model can also cope with reverberation by storing reverberated speech exemplars rather than clean speech exemplars.

5.2.2 Obtaining the exemplar weights

The non-negative exemplar weights $\mathbf{x}_{c,l}$ are obtained by minimizing the cost function,

$$d(\mathbf{y}_l, \mathbf{A}_{c,l}\mathbf{x}_{c,l}) + \sum_{m=1}^{M_{c,l}} x_{c,l}^m \Lambda_m \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \tag{5.2}$$

where $\mathbf{\Lambda}$ is an $M_{c,l}$-dimensional vector. The first term is the divergence between the observation vector and its approximation. The second term is a regularization term which penalizes the l_1-norm of the weight vector to produce a sparse solution. $\mathbf{\Lambda}$ contains non-negative values and controls how sparse the resulting vector $\mathbf{x}$ is. Defining $\mathbf{\Lambda}$ as a vector, the amount of sparsity enforced on different types of exemplars can be adjusted. Moreover, $\mathbf{\Lambda}$ values can be chosen depending on the SNR level for improved recognition performance.

The regularized optimization problem with the cost function in Equation (5.2) can be solved with various techniques including least absolute shrinkage and selection operator (LASSO) [163], approximate Bayesian compressive sensing [17], elastic net [192] and non-negative sparse coding (NSC) [79]. In this work, NSC is applied to obtain the exemplar weights that minimizes the cost function.

The approximation highly depends on the congruence of the representation of the speech and the divergence measure in Equation (5.2). Particularly, depending on the distribution of the speech and noise sources in the high-dimensional feature space, an appropriate divergence/distance measure has to be chosen in the sense that it weights the reconstruction error in each component (mel band) in a desired way.

The generalized Kullback-Leibler divergence (KLD) has been found to provide better results when used in conjunction with magnitude spectral features compared to the Euclidean distance in source separation, SR-based noise robust speech recognition and polyphonic music transcription [138, 151, 152, 162, 174]. Hence, we investigate the recognition performance of the proposed system using the generalized KLD for d. The generalized KLD is defined as

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k. \tag{5.3}$$

For NSC, we apply the relaxation and non-linear projection techniques proposed in [21] for faster convergence to the multiplicative update rule derived in [50] to minimize the cost function. The final multiplicative update rule is given by

$$\mathbf{x}_{c,l} \leftarrow (\mathbf{x}_{c,l} \odot ((\mathbf{A}_{c,l}^T(\mathbf{y}_l \oslash (\mathbf{A}_{c,l}\mathbf{x}_{c,l}))) \oslash (\mathbf{A}_{c,l}^T\mathbf{1} + \mathbf{\Lambda}))^{\cdot[\omega]})^{\cdot[1+\alpha]} \tag{5.4}$$

with $\odot$, $\oslash$ and $^{\cdot[\,]}$ denoting element-wise multiplication, element-wise division and element-wise exponentiation respectively. ω is a value between $(0, 2)$ and α is a very small positive number [21]. $\mathbf{1}$ is a Dl-dimensional vector with all elements equal to unity. Applying this update rule iteratively, the weight vector becomes sparse and the reconstruction error between the noisy speech vector and its approximation decreases monotonically.

5.2.3 Decoding

All observation matrices Y_l for $l_{\min} \leq l \leq l_{\max}$ are approximated using the combined dictionaries $\mathbf{A}_{c,l}$ of the corresponding length by applying the multiplicative update rule in Equation (5.4). To quantify the quality of the match, we use the reconstruction error between the noisy speech segments and their approximations. The first term of Equation (5.2) expresses the reconstruction error between a noisy speech segment of length l and its approximation.

The multiplicative update rule is applied iteratively until the reconstruction error provides enough discrimination between different classes. The number of iterations that satisfies this criterion has been investigated in pilot experiments. After a fixed number of iterations for all dictionaries, the reconstruction errors between the observation matrix Y_l and its approximations $\mathbf{A}_{c,l}\mathbf{x}_{c,l}$ are calculated for $l_{\min} \leq l \leq l_{\max}$. As the label of each dictionary is known, decoding is performed by applying a single-stage dynamic programming algorithm [128] to find the class sequence that minimizes the reconstruction error taking the grammar into account.

This kind of search problem can be visualized as a three-dimensional grid search over grid points (x, y, z) which are defined by the time frames x of a noisy speech segment, time frames y of its approximation and the dictionary number z [128]. Focusing on the noise robustness, noisy speech segments are only matched with the dictionaries of the same duration, i.e. no time warping is performed, within the scope of this study.

5.2.4 Dictionary normalization

The rows of the combined dictionaries (mel frequency bands) are scaled with the weights obtained as the squared sum of the mel frequency bands of training frame sequences to avoid the reconstruction error being dominated by a few bands only. The columns of the combined dictionaries (exemplars) are also l_p-normalized which has been shown to improve the recognition results in [187]. The same column normalization is applied to the observation matrices $\mathbf{Y}_l$.

The column normalization of the observation matrices replaces the exemplar start-up cost which is commonly used in exemplar matching-based systems in order to limit the number of exemplars that are used to explain the observed segments. Without the column normalization (or the exemplar start-up cost), the recognizer has the tendency to explain the observed segments using the shortest exemplars yielding unrealistic recognition results as they fit the data better due to the higher degree of freedom. l_p-normalization of the observation matrices scales the reconstruction errors by a factor that increases with the exemplar length, hence, the value of p can be tuned to balance the number of insertions and deletions. The proposed system applying l_p-normalization to the observation matrices has provided better recognition accuracies in the pilot experiments compared to adopting an exemplar start-up cost.

5.2.5 Compensating the silence scores

A known problem of sparse representation approaches working on magnitude spectra is that silences are hard to recognize: perfect silence is modeled with zero weights of all exemplars [50]. In a practical noisy mixture, silence is well-approximated by combining speech and noise exemplars with small weights, so all classes will score equally well. To overcome this problem, reconstruction errors of the dictionaries representing silence have to be compensated. For this purpose, we use a single overcomplete dictionary containing speech and noise exemplars to approximate the noisy speech and perform voice activity detection (VAD), i.e. predicting whether a noisy speech segment contains speech. Choosing an exemplar length L_s containing abundant samples from each class, we form a single dictionary by concatenating all speech exemplars from different classes plus noise exemplars that are extracted from the noise-only training sequences. After obtaining the exemplar weights for every noisy segment of length L_s, we reconstruct the speech components to detect the frames where speech activity exists. The schematic illustration of the single dictionary setup is given in Figure 5.2.

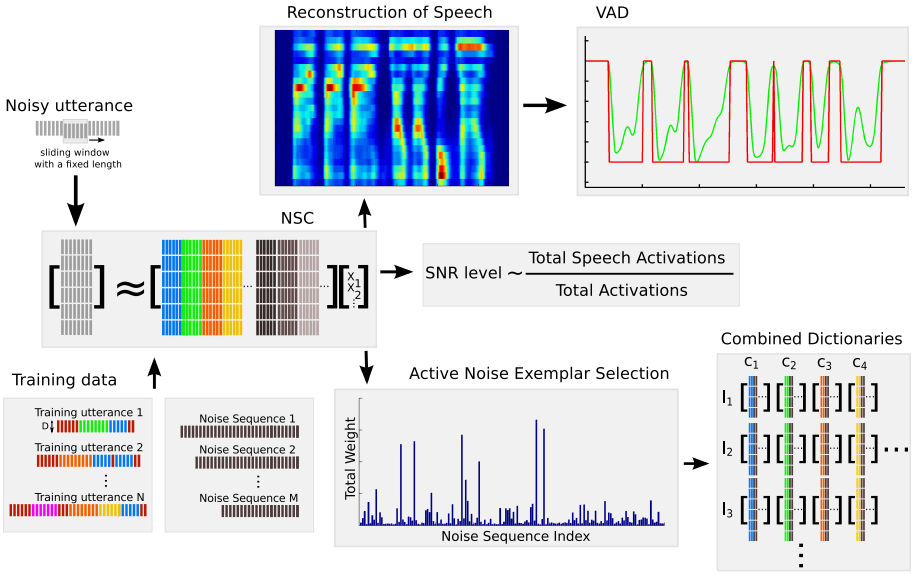


Figure 5.2: VAD, SNR estimation and Active Noise Exemplar Selection (ANES) - A single dictionary setup is proposed for VAD, SNR estimation and ANES. The speech weights are used to reconstruct the speech component providing information about the frames containing speech. SNR level is estimated as the ratio of the total speech weights to the total speech and noise weights in order to limit the estimation range to [0,1]. Finally, noise weights belonging to the noise exemplars that are extracted from the same noise sequence are accumulated to identify which noise sequences are able to model the actual noise conditions. Noise exemplars that are used in the recognition are extracted from the most active noise sequences.

The range of the reconstruction error values obtained for each class is SNR dependent. For high SNRs, speech exemplars get higher weights yielding a higher range in the reconstruction errors among different-class dictionaries of the same length. On the contrary, for low SNRs, the noise exemplars get higher weights resulting in very close reconstruction errors. To avoid overcompensation of the reconstruction errors at lower SNRs, we propose an SNR-dependent compensation factor. The SNR level is estimated as the ratio of the total speech weights to the total speech and noise weights. The details of the VAD and SNR estimation are given in Appendix A.1.

5.3 Dictionary Design

5.3.1 Motivation

Although exemplar-based modeling is known to provide better duration and trajectory modeling compared to the statistical models, a large amount of data is required to handle the acoustic variation among different utterances [30]. In order to reduce the high memory and computational power required for data handling, several exemplar selection algorithms are proposed for exemplar matching systems in [149, 158]. The main goal of these techniques is to remove less informative exemplars that are hardly used or whose presence result in inaccurate recognition and achieve comparable recognition accuracies using only a fraction of the exemplars. Moreover, several techniques have also been proposed to reduce the number of atoms organized in an overcomplete dictionary [53, 64, 86, 115].

As the speech exemplars are associated with a single speech unit, their length distribution is class-dependent which results in unevenly populated speech dictionaries. Speech dictionary design mainly involves increasing the number of exemplars in underpopulated speech dictionaries to avoid poor acoustic modeling or reducing the number of exemplars in highly populated dictionaries without a significant loss in the recognition accuracy. On the other hand, noise exemplars are extracted from noise-only training sequences for any arbitrary length. As a result, while there are a vast number of noise exemplars for every exemplar length, only the ones that match the actual noise conditions will be beneficial during recognition. Thus, noise dictionary design mainly focuses on accurate modeling of the background noise using the smallest possible number of noise exemplars.

Previously, we have described various speech exemplar selection techniques to limit the number of exemplars organized in undercomplete dictionaries [182]. The recognition experiments performed on clean speech have shown that using only 30% of the speech exemplars does not result in a significant loss in the recognition accuracy. Moreover, for the dictionaries that contain only a few speech exemplars, we apply *prewarping* to increase the number of exemplars by manipulating the same-class exemplars of different lengths [183].

In this section, we focus on the noise dictionary design and present the techniques that are used in the proposed framework. These dictionary design techniques are applied either to improve the acoustic modeling capabilities of the dictionaries and/or to reduce the dictionary sizes for less computational power and memory requirements.

5.3.2 Noise Dictionary Design

Active noise exemplar selection (ANES)

Previous experiments have shown that noise modeling using the same noise dictionaries for every noisy utterance provides very poor estimation of the noise source [181]. Using smaller noise dictionaries due to the computational restrictions results in inferior performance compared to the previous SR-based recognizers especially at lower SNR levels.

In this section, we introduce a design procedure for adaptive noise dictionaries which uses the noise weights that are provided by the single dictionary setup described in Section 5.2.5. This technique aims to select a small number of noise exemplars that can accurately model the actual noise conditions. An equal number of exemplars is extracted from numerous noise-only training sequences and stacked in a single noise dictionary. Before performing the recognition, the noisy utterance is approximated using the single dictionary containing these noise exemplars as shown in Figure 5.2. In order to identify which noise-only training sequences can accurately model the actual noise conditions, all weights belonging to the noise exemplars extracted from each noise-only training sequence are accumulated. With the same motivation as discussed in [182], noise exemplars used for the recognition are extracted from the most active noise-only training sequences, i.e. the ones with the highest weights.

Acquiring noise exemplars on the fly

Another technique that has been proposed for improved noise modeling in SR-based recognition systems is called *noise sniffing* [53]. This technique acquires noise exemplars on the fly from the immediate neighborhood of the target utterance. The extracted noise exemplars are added to the combined dictionaries and used for the recognition. In case of limited noise context, a small number of frames from the beginning and end of the target utterance are extracted and contained in combined dictionaries. Shifted copies of these frame sequences are also included to provide some degree of shift-invariance [52]. The VAD information extracted from the setup in Figure 5.2 is used to detect the speech onset and offset points.

SNR-dependent noise modeling

To find a compromise between the accuracy of the noise modeling and computational complexity, the amount of the noise exemplars in the combined

dictionaries is adjusted depending on the estimated SNR level. At lower SNRs, a larger number of noise-only training sequences are used for noise exemplar extraction. Consequently, computational complexity of the recognizer is reduced at high SNRs without loss of recognition accuracy while preserving the noise modeling capabilities at lower SNRs. Moreover, SNR-dependent noise modeling provides gains in the recognition accuracy of clean speech, as the dictionaries contain only a few noise exemplars during the recognition of clean speech.

5.4 Experimental Setup

5.4.1 Databases

CHIME-2

The small vocabulary track of the 2$^{\text{nd}}$ CHiME Challenge [172] addresses the problem of recognizing commands in a noisy living room. The clean utterances in the CHIME-2 data are taken from the GRID corpus [23] which contains utterances from 34 speakers reading 6-word sequences of the form *command-color-preposition-letter-digit-adverb*. There are 25 different letters, 10 different digits and 4 different alternatives for each of the other classes. Even though there is no silence between the words, leading silences of variable duration occur occasionally. The recognition accuracy of a system is calculated based on the correctly recognized letter and digit keywords.

The clean utterances are convolved with binaural room impulse responses with speaker head movement effects which are recorded in a living room. Then, the resulting reverberated utterances are mixed with binaural recordings of genuine room noise recorded in the same living room at SNR levels of 9, 6, 3, 0 ,-3 and -6 dB. The training set contains 500 utterances per speaker (17,000 utterances in total) with clean, reverberated and noisy versions. Noisy utterances are provided both in isolated or embedded form. Embedded recordings contain 5 seconds of background noise before and after the target utterance. The development and test sets contain 600 utterances from all speakers at each SNR level (3600 utterances in total for each set) both in isolated and embedded form. The immediate noise context of the target utterances is available in the embedded recordings. The development set also contains 600 noise-free reverberated utterances. All data has a sampling frequency of 16 kHz.

AURORA-2

The recognition performance of N-REM is further evaluated on the test set A and B of the AURORA-2 corpus [77]. The training material of AURORA-2 consists of a clean and a multi-condition training set, each containing 8440 utterances. The multi-condition training set was constructed by mixing the clean utterances with noise at SNR levels of 20, 15, 10 and 5 dB.

Test set A consists of 4 clean and 24 noisy datasets with four noise types (subway, babble, car and exhibition) at six SNR levels, 20, 15, 10, 5, 0 and -5 dB. The noise types of this test set match the multi-condition training set. Test set B has the same number of test sets with four different noise types (restaurant, street, airport, station) at the same SNR levels. Each subset contains 1001 utterances with one to seven digits 0-9 or oh. To reduce the simulation times, we subsampled the test sets by a factor of 4 (250 utterances per test set, 1000 utterances per SNR). All parameters are tuned on a different subset with 100 utterances from each test set. All data has a sampling frequency of 8 kHz.

5.4.2 Dictionary Creation and Implementation Details

The recognition system is implemented in MATLAB and GPUs are used to accelerate the evaluation of Equation (5.4). Two versions of N-REM have been investigated depending on the recognition tasks. The first version, which does not include the single dictionary setup, has been investigated on the CHIME-2 data as the silences between the words are assumed to be negligible. The second version including the single dictionary setup has been applied on the AURORA-2 task.

CHIME-2

The exemplars and noisy speech segments are represented as mel-scaled magnitude spectral features extracted with a 26 channel mel-scaled filter bank ($D = 26$). The frame length is 25 ms and the frame shift is 10 ms. The binaural data is averaged in the spectral domain to obtain 26-dimensional feature vectors.

The exemplars are extracted from the reverberated utterances in the training set according to the state-based segmentations obtained using the acoustic models in the toolkit provided with the database. Exemplars belonging to each speaker are organized in separate dictionary sets for speaker-dependent modeling yielding 34 different dictionary sets. Based on the availability of the

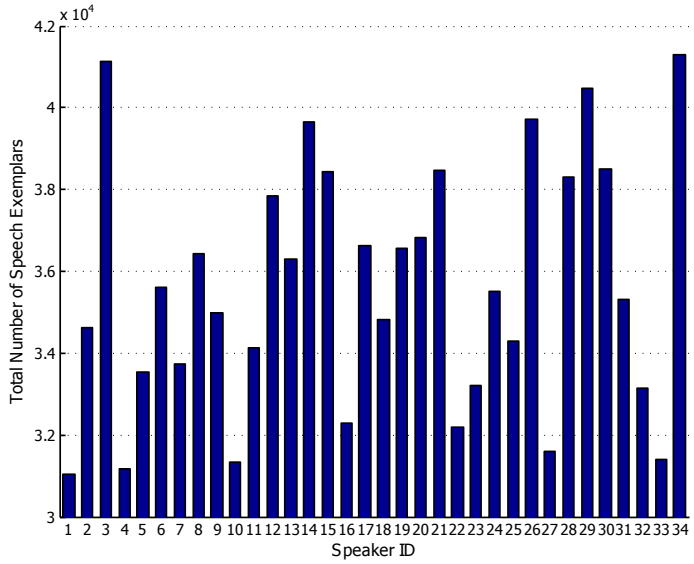


Figure 5.3: Number of speech exemplars for each speaker in the CHIME-2 Data

exemplars, the minimum and maximum exemplar lengths are 4 and 45 frames respectively.

After preliminary experiments, exemplars representing words turned out to provide poor acoustic modeling resulting in a high error rate. Half-word exemplars seemed to generalize sufficiently to unseen data. Half-word exemplars are extracted by cutting the word exemplars at the HMM state yielding the minimum average length difference between the two halves. Dictionary sizes vary with different classes and speakers. *Prewarping* [183] is applied to boost the modeling capabilities of the underpopulated speech dictionaries (especially for the ones belonging to letters due to the high number of alternatives and hence the small number of exemplars per class) and it is limited to a single frame. The number of exemplars in each dictionary after prewarping is limited to 50. The number of speech exemplars for each speaker after prewarping is shown in Figure 5.3.

The silences between the words are assumed to be negligible, hence, dictionaries representing a silence class are not used. This comes with several advantages as the silence compensation discussed in Section 5.2.5 is not a requirement. However, the isolated utterances in the training, development and test sets occasionally contain leading silence of variable duration. To overcome the mismatches in silence duration during the decoding, the number of frames

belonging to the first HMM state of the first word in the reverberated training data is limited to 10 frames while extracting the exemplars. Furthermore, during recognition, the decoding is repeated 5 times each time omitting 5 frames from the beginning. The class sequence yielding the minimum reconstruction error per frame is then chosen to be the recognition output.

The noise dictionaries used in the experiments contain 400 noise exemplars that are acquired on the fly (cf. Section 5.3.2) from the immediate neighborhood of the target utterance in both directions until the frames belonging to other target utterances.

This recognizer uses SNR-independent $\mathbf{\Lambda}$ values in Equation (5.4). Elements of $\mathbf{\Lambda}$ in Equation (5.4) are tuned for the highest recognition accuracy on the development data and set to 1.75 and 3 for speech and noise exemplars respectively. The multiplicative update rule is iterated 25 times to obtain the exemplar weights. ω and α are set to 1.75 and 0.008 respectively. The columns of the combined dictionaries and observation matrices are l_2-normalized.

AURORA-2

The recognition experiments performed on AURORA-2 data are organized in two parts. In the first part, the performance of the ANES technique has been evaluated on both test sets at the SNRs of -5, 0 and 5 dB. The recognition accuracies obtained using adaptive dictionaries are compared to the ones obtained using the fixed noise dictionaries [181]. In the second part, the recognition performance of N-REM with the adaptive noise dictionaries is compared to the other state-of-the-art recognizers.

The speech exemplars are extracted from the clean training set of AURORA-2 database [77] which contains 8440 utterances with one to seven digits in American English. Acoustic feature vectors are represented in mel-scaled magnitude spectra with 23 frequency bands. The speech exemplars representing half-digits are segmented by a conventional HMM-based system. There are in total 52,305 speech exemplars excluding 990 silence exemplars. The minimum and maximum exemplar lengths are 8 and 40 frames respectively. The number of speech exemplars extracted for each length and class is shown in Figure 5.4. Exemplars longer than 40 frames are omitted to limit the number of dictionaries.

In the first part of the experiments, the oracle VAD, a single SNR estimate and $\mathbf{\Lambda}$ value for speech and noise exemplars are used at each SNR level, in order to control the impact of irrelevant parameters on the recognition accuracy. For this purpose, VAD is obtained by applying an external energy-based VAD detector to the clean versions of the noisy utterances and SNR estimates are set

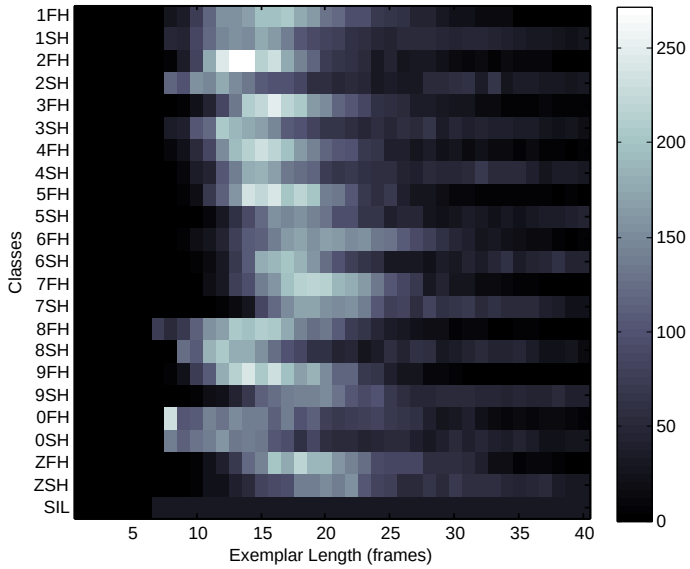


Figure 5.4: Exemplar length distribution in the AURORA-2 database (The classes are half-digits, e.g. '5FH' stands for the first half of digit '5'. 'O', 'Z' and 'SIL' stands for 'oh', 'zero' and 'silence' respectively. The bar on right gives the range of the counts.)

to 0.05, 0.15 and 0.25 for SNRs of -5, 0 and 5 dB respectively. The $\mathbf{\Lambda}$ values for speech exemplars are set to 0.5, 1 and 2 for SNRs of -5, 0 and 5 dB respectively. The $\mathbf{\Lambda}$ values for noise exemplars are set to half of the speech values.

The fixed dictionaries are extracted from 16 longest noise-only training sequences (1 sequence from each set) with jumps of 4 frames. The noise-only training sequences are obtained by removing speech from the noisy utterances in the multi-condition training set. As a result, the fixed dictionaries contain between 547-589 noise exemplars depending on the frame length.

Adaptive dictionaries are created based on the noise weights that are obtained using the single dictionary setup. The single dictionary contains noise exemplars that are extracted from either 160 or 800 longest noise-only training sequences (10 or 50 sequences from each set). From each noise-only training sequence, 10 noise exemplars are extracted with equal jumps resulting in 1600 or 8000 noise exemplars. The single dictionary also contains 2200 speech exemplars (100 exemplars from each class excluding silence). It uses speech and noise exemplars containing 15 frames. First and last 20 frames of the target utterances are

assumed not to contain speech and 150 noise exemplars with 15 frames (5 exemplars and 70 shifted copies from each end) are extracted as described in Section 5.3.2 and concatenated to the single dictionary. The speech and noise weights are obtained after 300 iterations. Elements of $\mathbf{\Lambda}$ in Equation (5.4) are set to 2 and 1.75 for speech and noise exemplars respectively.

For both fixed and adaptive dictionaries, *noise sniffing* is also performed at the recognition phase. The way *noise sniffing* is performed is slightly different than in the single dictionary setup. Based on the VAD, the speech onset and offset frames are detected and the number of sniffed frames is increased if these points are beyond 20 frames. For each exemplar length l, first l frames and $l-1$ shifted copies are added to the combined dictionaries. Sniffed frame sequences are linearly interpolated for the exemplar lengths that are larger than the number of sniffed frames. After concatenating the different speech and noise dictionaries, the system ends up containing 675 dictionaries of 23 different classes (half-digits plus silence). The combined dictionaries and observation matrices are l_3-normalized to balance the deletions and insertions for all SNR levels. The multiplicative update rule is iterated 100 times for convergence of all frame lengths with $\omega = 1.75$ and $\alpha = 0.008$.

In the second part of the experiments, we compare the recognition performances of N-REM using the single dictionary containing noise exemplars from 800 noisy-only training sequences with other recognizers. The details given in the initial part are the same, except that the single dictionary setup is also used for the VAD, SNR estimation. SNR-dependent noise modeling is performed to reduce the computational load for higher SNR levels (cf. Section 5.3.2). The details of the SNR-dependent ANES technique are given in Appendix A.2. Recognition is performed after obtaining the adaptive noise dictionaries, VAD and SNR estimate. $\mathbf{\Lambda}$ values in Equation (5.4) are SNR-dependent with a ratio of 0.3 between noise and speech exemplars. $\mathbf{\Lambda}$ for speech exemplars are set to a scalar multiple c of the SNR_{est} which is defined in Appendix A.1. c is set to 8. Maximum values of $\mathbf{\Lambda}$ for speech and noise exemplars are set to 8 and 2.4 respectively.

5.4.3 Evaluation Metrics

We have opted for the metrics which have been traditionally used for the evaluation of the databases described in Section 5.4.1 for comparability with the previous literature. The keyword recognition accuracy (RA) is used to evaluate the system performance on the CHIME-2 data. The word error rate (WER) has been used to quantify the recognition accuracy for the AURORA-2 digit recognition task.

5.5 Results and Discussion

The recognition performance of N-REM is compared with a standard GMM recognizer and two noise robust SR-based recognition techniques using fixed-length exemplars in a single overcomplete dictionary, namely sparse classification (SC) and feature enhancement (FE) [50].

For CHIME-2 data, the GMM recognizer uses speaker-dependent acoustic models trained on noisy data. These results are obtained using the HTK recognition toolkit and the details are available at the 2nd CHIME Challenge website[1]. The details of the FE and SC recognition systems such as feature extraction schemes and dictionary sizes are described in [51]. The FE recognition system refers to the baseline NMF system trained on the reverberated data in [51].

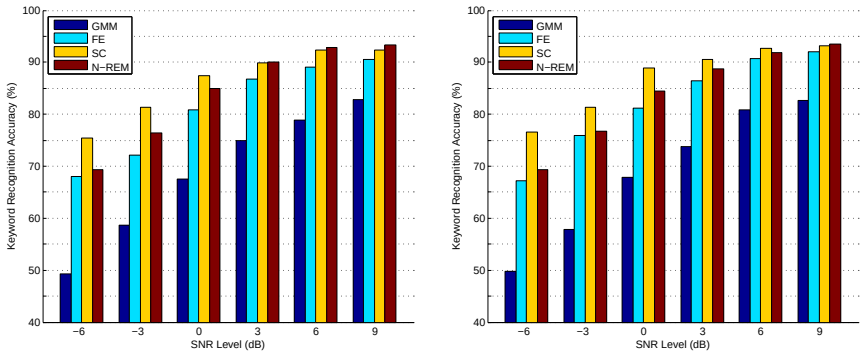
The GMM and other SR-based recognizers applied on AURORA-2 database are detailed in [52]. The SC and FE recognizers achieve among the best known results on AURORA-2, especially at lower SNRs, performing significantly better than for instance the ETSI advanced front-end (AFE) which has been considered as a reference for the AURORA-2 database [78]. The GMM and FE recognition systems are trained on the multi-condition training set. The FE and SC recognition systems use fixed-length exemplars containing 30 frames. We have performed recognition experiments on the same subset containing 1000 utterances from each SNR to obtain comparable recognition results.

CHIME-2 Recognition Experiments

The keyword recognition accuracies obtained on the development and test sets of the CHIME-2 data are given in Figure 5.5. The recognition performances on the development and test sets are similar for all systems. The SC recognizer performs better especially at lower SNRs compared to the other recognizers providing recognition accuracies of 76.5% and 81.3% at -6 and -3 dB on the test set respectively. N-REM yields recognition accuracies of 69.3% and 76.4% at the same SNR levels which is slightly higher than 68.0% and 75.9% of the FE recognizer.

At higher SNRs, N-REM provides comparable results with the SC recognizer. The recognition accuracies obtained with N-REM at 6 and 9 dB are 91.9% and 93.5% compared to 92.7% and 93.2% of the SC recognizer. The FE recognizer performs slightly worse than these recognizers with recognition accuracies of

[1] `http://spandh.dcs.shef.ac.uk/chime_challenge/chime2013/chime2_task1.html`



(a) Development Set (b) Test Set

Figure 5.5: CHIME-2 Recognition Results - The recognition results obtained using N-REM and the other recognizers (GMM, FE, SC) are given for SNR levels from -6 to 9 dB on both the development and test set.

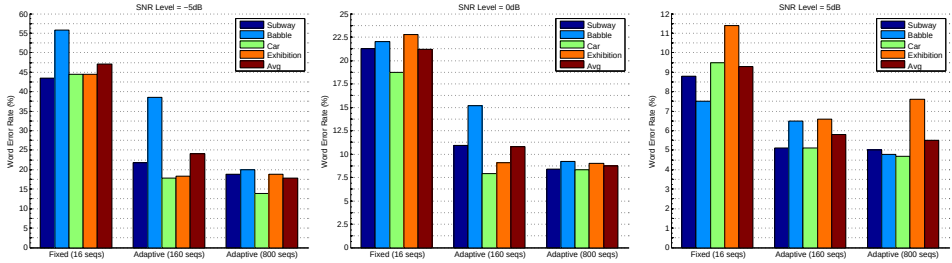
90.7% and 92.0% at same SNR levels. The GMM recognizer trained on the noisy data provides substantially worse results at all SNR levels.

From these results, it can be concluded that N-REM provides a sufficient level of noise robustness using only a small set of noise exemplars on condition that they accurately capture the spectro-temporal properties of the non-stationary noise corrupting the target utterance. Compared to the other SR-based systems, a much lower number of iterations is required for the reconstruction error to provide enough discrimination between classes thanks to the competition among the compact class-dependent dictionaries.

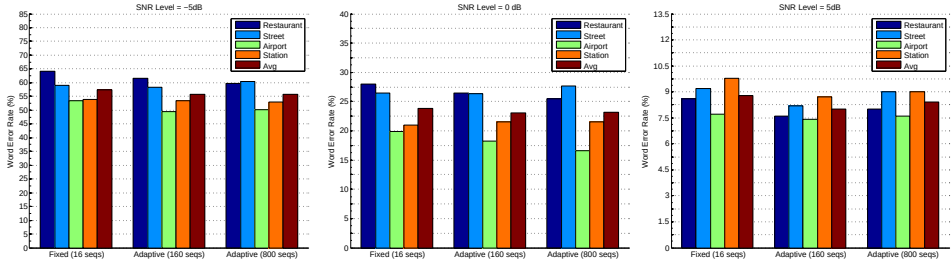
Separation accuracy of speech and noise sources highly depends on the exemplar length. In [50], it has been shown that using longer exemplar sizes provides better separation of speech and noise. Even though the proposed approach comes with flexibility of using multiple length exemplars, the duration distribution of the classes in the training data has an impact on the separation performance. For instance, the exemplars representing half-letters mostly contain 4-10 frames which makes the recognition less robust to noise compared to the SC and FE techniques using fixed-length exemplars of 20 frames at lower SNRs.

Finally, N-REM suffers from the lack of training data compared to the other recognition systems using GMMs or fixed-length exemplars, as the exemplars belonging to each class are distributed among dictionaries of multiple lengths resulting in underpopulated dictionaries.

Noise Dictionary Design with ANES



(a) Test Set A



(b) Test Set B

Figure 5.6: Comparison of the recognition results using fixed and adaptive noise dictionaries on AURORA-2. The upper half of the figure presents the recognition results performed on test set A at SNR levels -5, 0 and 5 dB. On the left, the results obtained with the fixed noise dictionaries are provided. In the middle and right, the results yielded by adaptive dictionaries using either 160 or 800 noise-only training sequences are given. In each graph, results obtained on each noise type are given separately and the fifth bar of each experiment is the mean of all noise types. The lower half presents the results obtained on test set B at the same SNR levels.

We have performed recognition experiments on the test set A and B of the AURORA-2 database to compare the performance of the adaptive noise dictionaries obtained using the ANES technique compared to the use of fixed noise dictionaries. The recognition results are given in Figure 5.6. The upper half of the figure presents the results on test set A at -5, 0 and 5 dB. The lower half presents the results on test set B at the same SNR levels. Results obtained on each noise type are given separately and the fifth bar of each experiment is the mean of all noise types.

In the case of matched noise, using adaptive dictionaries provides large

improvements at all SNR levels. The proposed recognizer with ANES that selects noise exemplars from 160 noise-only training sequences yields a word error rate (WER) of 24.1% compared to the 47.1% of the fixed dictionaries. Increasing the number of noise-only training sequences from 160 to 800 provides an absolute improvement of 6.2% reducing the WER to 17.9%. The largest improvement is obtained in case of babble noise with a decrease in the WER from 38.5% to 20.0%.

Using 800 training sequences in the final recognizer seems to be a reasonable choice as the WERs on different noise types tend to converge. At 0 dB and 5 dB, similar improvements are obtained with WERs of 21.2% and 9.3% using the fixed dictionaries compared to 8.7% and 5.5% using the ANES technique with 800 training sequences respectively.

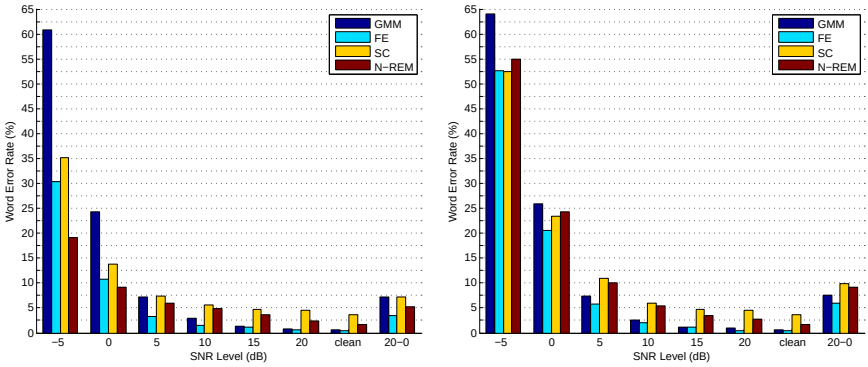
The results on test set B have shown that the ANES technique yields marginal improvements in the case of mismatch noise. At -5 dB, the adaptive dictionaries with 160 training sequences provide a WER of 55.7% compared to 57.5% of the fixed dictionaries with an absolute improvement of 1.8%. At 0 dB and 5 dB, the WERs are reduced from 23.8% and 8.8% to 23.1% and 8.0% respectively. Increasing the number of training sequences does not improve the performance further.

Considering these results, the proposed way of noise modeling was found to be very effective on matched noise yielding significant improvements in the recognition at lower SNRs. However, the improvement obtained on the mismatched noise is limited as the ANES technique cannot find noise-only training sequences that can accurately model the noise.

AURORA-2 Recognition Experiments

Finally, we compare the performance of the N-REM recognizer using adaptive noise dictionaries with the aforementioned recognizers. The WERs obtained on the test set A and B of the AURORA-2 database are given in Figure 5.7. The results on test set A are given on the left side of the figure. N-REM performs better at -5 dB and 0 dB with WERs of 19.1% and 9.2% compared to 30.4% and 10.7% of the FE and 35.2% and 13.8% of the SC recognition systems respectively. This demonstrates the effectiveness of the ANES technique on the matching noise scenarios. At these SNR levels, the GMM recognizer performs considerably worse than the SR-based methods.

At higher SNRs, the performance of FE and GMM recognizers is better than N-REM and SC, thanks to the GMM-based back-end used in conjunction with MFCC features. N-REM performs slightly better than SC with WERs of



(a) Test Set A

(b) Test Set B

Figure 5.7: AURORA-2 Recognition Results - The recognition results obtained using N-REM and the other recognizers (GMM, FE, SC) are given for SNR levels from -5 to 20 dB on the same subset of test set A and B containing 1000 utterances per SNR level. The average WERs for the SNR levels between 20 and 0 dB are given on the rightmost bar of each figure.

4.9% and 5.6% at 10 dB, 3.6% and 4.8% at 15 dB, and 2.4% and 4.5% at 20 dB respectively. The WER provided by N-REM on the clean speech is 1.7% compared to 0.7% of the GMM, 0.5 of the FE and 3.6% of the SC recognizer.

On the right side of the same figure, the WERs on test set B are presented. N-REM performs slightly worse than the other SR-based recognizers with WERs of 55.0% and 24.3% at -5 and 0 dB. These results are worse than the results provided by the SC with WERs of 52.4% and 23.5% and FE with the WERs of 52.6% and 20.5% for the same SNRs respectively. GMM performs worse than the SR-based recognizers. At higher SNRs, the results follow a similar trend as the results on test set A.

5.6 General Discussion

5.6.1 Speech Recognition Performance

The main goal of this work was to propose a novel exemplar-based speech recognition framework, N-REM. Unlike conventional exemplar matching approaches, it is noise robust by explicitly modeling noise and unlike the

recently proposed exemplar-based noise robustness techniques it builds on, its exemplars model speech units rather than arbitrary fixed-length exemplars.

While the results show that the proposed framework is quite noise robust, evaluations on the clean speech of AURORA-2 do show that N-REM does not perform as well as a GMM/HMM-based systems. While N-REM does perform better than SC, the other purely exemplar-based recognizer in our comparisons, it is hampered by the same constraints as SC: the necessity of working with a feature representation for which speech and noise are (approximately) additive. Combination of the generalized KLD and the mel-scaled magnitude spectral features are not as discriminative as the complex GMM distributions that are used in conjunction with MFCC features.

At lower SNRs, the proposed framework performs reasonably well - at SNRs 10 to 0 dB, N-REM performs comparably to the SC and FE systems on both CHIME-2 and AURORA-2 data. At even lower SNRs, N-REM outperforms FE on CHIME-2 and both FE and SC on test set A of AURORA-2. It performs comparably to SC and FE on test set B of AURORA-2. In fact, to the best of our knowledge, the WER of 19.1% obtained by N-REM at -5 dB is the best result ever reported on test set A of AURORA-2.

The exemplar-based techniques FE, SC and N-REM have in common that there is a performance gap between matched and mismatched noise cases. As such, the use of exemplars is most applicable in scenarios where the expected noise types can be predicted (and thus stored in the noise dictionary), or when matching exemplars can be readily obtained from the environment, a scenario which is mimicked in CHIME-2 data. For N-REM, this gap is larger than for SC and FE since N-REM uses smaller noise dictionaries. This results in a lower probability of having a suitable noise exemplar in the combined dictionaries with a similar spectral content with the unseen noise types.

At the same time, the SNR-dependent noise modeling and the proposed noise dictionary design technique ANES are very effective at picking matching noise exemplars from training material, especially for traditionally difficult noise types such as babble noise. Depending on the memory and computational limitations, the size of the noise repository can be easily increased to have better coverage of multiple noise types, and thus improve the performance.

5.6.2 Computational Effort

The computational bottleneck of the proposed framework is the multiplicative update rule given in Equation 5.4. In practice, the simulation of the proposed technique has benefited substantially from the use of GPUs. All recognition

experiments have been performed using an NVIDIA Tesla C2070 GPU. To quantify the simulation times for each task, we have timed the recognition processes in MATLAB for each utterance and averaged the simulation time per utterance over each test set. On the CHIME-2 data, the average recognition time is obtained by averaging the recognition time over 600 test utterances. For these utterances, the mean duration is 1.8 seconds and the average recognition time is 26.1 seconds with a standard deviation of 2.7 seconds.

The SNR-dependent noise dictionaries yield different simulation times at each SNR level in AURORA-2 database. At -5 dB, the recognizer uses the highest number of noise exemplars, hence, the longest simulation times are expected at this SNR level. After averaging over a set containing 250 utterances with a mean duration of 1.7 seconds, the average recognition time is found to be 42.5 seconds with a standard deviation of 15.2 seconds. For higher SNRs, the SNR-dependent noise modeling has reduced the simulation times as the combined dictionaries contain less noise exemplars. On clean speech, the average recognition time reduces to 22.5 seconds with a standard deviation of 8.2 seconds. These average recognition times also include the time required for the single dictionary setup, which varies between 3-5 seconds depending on the utterance length.

5.7 Conclusion

In this chapter, we have introduced a novel recognition framework (N-REM) which performs noise robust speech recognition using multiple-length exemplars associated with a single speech unit. For each length, these speech exemplars are organized in separate speech dictionaries and they are concatenated with a noise dictionary forming the combined dictionaries that can model speech and noise mixtures. Using the combined dictionaries, noisy speech segments are approximated as a linear combination of the exemplars. The decoding is performed based on the quality of the match quantified by the reconstruction error between the noisy speech segments and their approximations.

Several design techniques are applied to noise dictionaries to have effective noise modeling with a small amount of noise exemplars. Firstly, we have introduced the *active noise exemplar selection* (ANES) technique which extracts noise exemplars from the training noise-only sequences that get high weights obtained using the single dictionary. *Noise sniffing* is applied to extract exemplars from the immediate noise context of the target utterance. Finally, SNR-dependent noise modeling is adopted in order to find a compromise between the noise modeling accuracy and computational restrictions.

We have performed several recognition experiments on the small vocabulary track of the 2nd CHiME Challenge and AURORA-2 database to investigate the noise robustness of the proposed framework. Initially, we have compared the performance of the adaptive noise dictionaries obtained using the ANES technique with the use of fixed noise dictionaries. The recognition results have shown that N-REM using adaptive noise dictionaries yields substantially higher recognition accuracies at lower SNRs. Moreover, we have compared the recognition performance of N-REM using the adaptive noise dictionaries with the other GMM and SR-based recognizers. N-REM using all aforementioned noise dictionary design techniques provides a higher degree of noise robustness in case of matched noise on the AURORA-2 database achieving WERs of 19.1% and 9.2% at SNR levels of -5 and 0 dB respectively. At higher SNRs, FE and GMM recognizers perform better than N-REM and SC thanks to the GMM/HMM back-end with MFCC features.

Chapter 6

Noise Dictionary Design for N-REM

This chapter investigates an adaptive noise dictionary design approach to achieve an effective and computationally feasible noise modeling for the noise robust exemplar matching (N-REM) framework. N-REM approximates noisy speech segments as a linear combination of multiple length exemplars in a sparse representation (SR) formulation. Compared to the previous SR techniques with a single overcomplete dictionary, N-REM uses smaller dictionaries containing considerably fewer noise exemplars. Hence, the noise exemplars have to be selected with care to accurately model the spectrotemporal content of the actual noise conditions. For this purpose, in a previous work, we introduced a noise exemplar selection stage before performing recognition which extracts noise exemplars from a few noise-only training sequences chosen for each target noisy utterance. In this work, we explore the impact of the several design parameters on the recognition accuracy by evaluating the system performance on the CHIME-2 and AURORA-2 databases.

This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Adaptive Noise Dictionary Design for Noise Robust Exemplar Matching of Speech*", Submitted to EUSIPCO 2015.

6.1 Introduction

Using exemplars in a sparse representation (SR) formulation to model noisy speech has provided major improvements in the automatic speech recognition (ASR) performance compared to conventional approaches such as hidden Markov models (HMM) under adverse conditions [143]. Previously, we have proposed an ASR system that performs noise robust exemplar matching (N-REM) [184] using exemplars of multiple lengths, each associated with a single speech unit such as phones, syllables, half-words or words similar to [29]. Exemplars of different length are organized in separate dictionaries based on the associated speech unit (class) and length unlike the previous SR-based systems [50,92,138] using a single dictionary with fixed-length exemplars. Using separate dictionaries for each class provides better classification as input speech segments are approximated as a linear combination of exemplars belonging to the same class only [181].

The N-REM dictionaries are substantially less populated compared to a single overcomplete dictionary, as the speech exemplars are associated with a single speech unit and their length distribution is class-dependent which results in unevenly populated speech dictionaries. Unlike the speech exemplars, noise exemplars are extracted from noise-only training sequences for any arbitrary length. As a result, while there are a large number of available noise exemplars for each exemplar length, only the ones that match the actual test noise conditions will be essential for accurate recognition. Thus, noise dictionary design mainly focuses on accurate modeling of the background noise using the smallest possible number of noise exemplars. Previous experiments have shown that rudimentary noise modeling approaches, e.g. using *fixed* noise dictionaries, provide very poor estimation of the noise source [181]. Using much smaller noise dictionaries due to computational restrictions compared to the previous SR-based recognizers with fixed-length exemplars results in inferior performance especially at lower SNR levels. For this reason, we have proposed an adaptive noise exemplar selection technique which chooses the best matching noise-only training sequences from a noise repository using a selection dictionary and extracts the noise exemplars that are used during the recognition from these sequences [184]. In this chapter, we further explore the impact of several design parameters, e.g. size of the noise repository and the amount of selected noise exemplars, on the recognition performance to reach a compromise between the noise robustness of the recognizer and the computational complexity.

6.2 Noise robust exemplar matching

Training frame sequences representing various noise-free speech units (speech exemplars), each comprised of D mel bands and spanning l frames, are extracted from the alignments obtained with an HMM-based recognizer and reshaped into a single vector and stored in the columns of a speech dictionary $\mathbf{S}_{c,l}$: one for each class c and each length l. Similarly, a single noise dictionary $\mathbf{N}_l$ for each length l is formed by reshaping noise exemplars. Each speech dictionary is concatenated with the noise dictionary of the same length to form a combined dictionary $\mathbf{A}_{c,l} = [\mathbf{S}_{c,l}\, \mathbf{N}_l]$ of dimensionality $(D \cdot l) \times M_{c,l}$ where $M_{c,l}$ is the total number of speech and noise exemplars.

An observed noisy speech segment of length T frames is also reshaped into vectors by applying a sliding window approach [50] with window length of l frames and stored in an observation matrix $\mathbf{Y}_l = [\mathbf{y}_l^1, \mathbf{y}_l^2 ..., \mathbf{y}_l^{(T-l+1)}]$ of dimensionality $(D \cdot l) \times (T - l + 1)$ for $l_{\min} \leq l \leq l_{\max}$ where $l_{\min}$ and $l_{\max}$ are the smallest and largest speech exemplar lengths respectively. For every class c, each observation vector $\mathbf{y}_l$ is expressed as a linear combination of the exemplars that are stored in the dictionaries of the same length, $\mathbf{y}_l \approx \mathbf{A}_{c,l}\mathbf{x}_{c,l}$ for $x_{c,l}^m \geq 0$ where $\mathbf{x}_{c,l}$ is an $M_{c,l}$-dimensional non-negative weight vector. The exemplar weights are obtained by minimizing the cost function $d(\mathbf{y}_l, \mathbf{A}_{c,l}\mathbf{x}_{c,l}) + \sum_{m=1}^{M_{c,l}} x_{c,l}^m \Lambda_m$ for $x_{c,l}^m \geq 0$ where $\mathbf{\Lambda}$ is an $M_{c,l}$-dimensional vector which contains non-negative values and controls how sparse the resulting vector $\mathbf{x}$ is. The generalized Kullback-Leibler divergence (KLD) is used for d which is commonly used in source separation problems and shown to produce better results than Euclidean distance when used in conjunction with mel-scaled spectral features [174]. The generalized KLD is defined as $d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k$.

The regularized optimization problem with the aforementioned cost function is solved with non-negative sparse coding (NSC) [79]. For NSC, we apply the multiplicative update rule given in [184] to obtain the exemplar weights. In practice, all observation matrices $\mathbf{Y}_l$ for $l_{\min} \leq l \leq l_{\max}$ are approximated using the combined dictionaries $\mathbf{A}_{c,l}$ of the corresponding length by applying the multiplicative update rule. After a fixed number of iterations, the reconstruction errors between each observation matrix $\mathbf{Y}_l$ and its approximation are calculated. As the label of each dictionary is known, decoding is performed by applying dynamic programming to find the class sequence that minimizes the reconstruction error.

6.3 Selection Dictionary Design

The accuracy of the noise modeling depends on the congruence of the spectrotemporal content of the noise exemplars and actual noise conditions contaminating the target utterance. Therefore, for each noisy utterance, a few noise-only training sequences that are able to model the background noise are selected on the fly and the noise exemplars in $\mathbf{N}_l$ are extracted from these sequences. The selection is performed by applying NSC using a selection dictionary $\mathbf{A}^*_{L_s} = [\mathbf{S}^*_{L_s}\ \mathbf{N}^*_{L_s}]$ containing speech exemplars from all classes and noise exemplars from different noise-only training sequences. The superscript * marks the dictionaries used in the noise exemplar selection. The speech dictionary $\mathbf{S}^*_{L_s}$ is obtained by concatenating an equal number of speech exemplars of the same length from each class. The length L_s can be set to any exemplar length containing abundant speech exemplars from each class.

For the noise dictionary $\mathbf{N}^*_{L_s}$, a noise repository of F noise-only training sequences is created and G noise exemplars are extracted from each noise-only training sequence with an equal frame shift. In total, $\mathbf{N}^*_{L_s}$ contains $F \cdot G$ noise exemplars. Once the selection dictionary $\mathbf{A}^*_{L_s}$ is formed, the observation matrix $\mathbf{Y}_{L_s}$ of length L_s is approximated as a linear combination of the exemplars in the selection dictionary $\mathbf{Y}_{L_s} \approx \mathbf{A}^*_{L_s}\mathbf{x}_{L_s}$ for $\mathbf{x}_{L_s} \geq 0$. By accumulating the weights of all noise exemplars extracted from the same training sequence, a total weight for each training sequence is obtained. Evidently, the training sequences having higher weights are expected to model the spectrotemporal properties of the background noise [182]. Hence, the noise dictionaries $\mathbf{N}_l$ for $l_\mathrm{min} \leq l \leq l_\mathrm{max}$ that are used during the recognition contain noise exemplars extracted from X training sequences with the highest weights.

Noise sniffing [53] is also applied for acquiring noise exemplars on the fly from the immediate neighborhood of the target utterance. The extracted noise exemplars are contained in the noise dictionaries, i.e., $\mathbf{N}^*_{L_s}$ as a part of the selection dictionary. Shifted copies of these frame sequences are also included to provide some degree of shift-invariance [52].

6.4 Experimental setup

6.4.1 Databases

The training material of AURORA-2 [77] consists of a clean and a multi-condition training set, each containing 8440 utterances. The multi-condition training set was constructed by mixing the clean utterances with noise at SNR levels of 20,

15, 10 and 5 dB. Test set A and B consist of 4 clean and 24 noisy datasets at six SNR levels between -5 and 20 dB. The noise types of test set A match the multi-condition training set. Each subset contains 1001 utterances with one to seven digits 0-9 or oh. To reduce the simulation times, we subsampled the test sets by a factor of 4 (1000 utterances per SNR).

The small vocabulary track of the 2nd CHiME Challenge [172] addresses the problem of recognizing commands in a noisy and reverberant living room. The clean utterances contain utterances from 34 speakers reading 6-word sequences of the form *command-color-preposition-letter-digit-adverb*. There are 25 different letters, 10 different digits and 4 different alternatives for each of the other classes. The recognition accuracy of a system is calculated based on the correctly recognized letter and digit keywords.

6.4.2 Exemplar extraction and implementation details

The speech exemplars are extracted from the clean training set of AURORA-2 data. Acoustic feature vectors are represented in mel-scaled magnitude spectra with 23 mel bands. The speech exemplars representing half-digits are segmented by a conventional HMM-based system. There are in total 52,305 speech exemplars excluding 990 silence exemplars. The minimum and maximum exemplar lengths are 8 and 40 frames respectively. Exemplars longer than 40 frames are omitted to limit the number of dictionaries. The noise-only training sequences are obtained by removing speech from the noisy utterances in the multi-condition training set. The *fixed* noise dictionaries are extracted from the 16 longest noise-only training sequences with shifts of 4 frames. Consequently, the fixed dictionaries contain between 547-589 noise exemplars depending on the exemplar length. The selection dictionary contains noise exemplars that are extracted from the longest noise-only training sequences. The amount of noise exemplars in the selection dictionaries depends on the chosen F and G value. The selection dictionary also contains 2200 speech exemplars. It uses speech and noise exemplars containing 15 frames. For AURORA-2, an SNR-dependent X value is used as it provides an improved recognition accuracy and reduced computational load at higher SNR levels by using less noise exemplars. The number of noise exemplars extracted from each sequence varies between 77 and 170. The further details of the SNR-dependent noise modeling is given in [184]. The word error rate is used to quantify the recognition accuracy on AURORA-2 data.

The exemplars and noisy speech segments of CHIME-2 data are represented as mel-scaled magnitude spectral features extracted with a 26 channel mel-scaled filter bank ($D = 26$). The frame length is 25 ms and the frame shift is 10 ms.

The binaural data is averaged in the spectral domain to obtain 26-dimensional feature vectors. Half-word exemplars belonging to each speaker are organized in separate dictionary sets for speaker-dependent modeling yielding 34 different dictionary sets. Based on the availability of the exemplars, the minimum and maximum exemplar lengths are 4 and 40 frames respectively. The baseline system performs recognition using noise dictionaries containing 400 sniffed noise exemplars. Each embedded utterance in the development and test set is segmented into noise-only sequences by removing all target utterances. G=5 noise exemplars of 25 frames are extracted from each noise-only sequence and stored in the single noise dictionary. The single noise dictionary size varies depending on the number of available noise-only sequences for each embedded recording. The adaptive noise modeling only evaluates the noise-only sequences that are extracted from the embedded recording which contains the target utterance. The number of noise exemplars extracted from each sequence varies between 95 and 195. The single speech dictionary contains 2354 full-word exemplars (maximum 50 exemplars from 51 classes) of 25 frames. The full-word exemplars are used in the single speech dictionary, as there is no exemplar length L_s containing a vast number of samples from each half-word class. The keyword recognition accuracy is used to evaluate the system performance on the CHIME-2 data.

6.5 Results and discussion

The recognition experiments performed on AURORA-2 data investigate the influence of the selection dictionary size, i.e. the noise repository size F and the number of exemplars extracted from each training sequence in the repository G, on the recognition performance. Choosing an SNR-dependent X best matching training sequences for the recognition is kept in the AURORA-2 experiments [184]. For CHIME-2 data, the selection dictionary is extracted from the noise-only segments of each embedded sequence which results in a fixed value of F. Hence, the CHIME-2 experiments investigate different settings of forming the noise dictionaries using the adaptive noise modeling approach and/or noise sniffing by varying the value of X. For this purpose, we compare the baseline recognizer using only sniffed exemplars with novel systems adopting adaptive noise modeling with and without the sniffed exemplars.

The performance of the adaptive noise modeling has been evaluated on both test sets of AURORA-2 data at the SNRs of -5, 0 and 5 dB and the results are presented in Table 6.1. The best results of the proposed setup are given in bold. The details of the other recognition systems can be found in [52]. In these recognition experiments, we compare the word error rates (WER) obtained

Table 6.1: Word error rates in percentages obtained on test set A and B of the AURORA-2 data

(a) Test set A

SNR(dB)	**-5**			**0**			**5**		
NREM (*Fixed*)	47.1			21.2			9.3		
GMM/HMM	60.8			24.3			7.3		
SC	35.2			13.8			7.4		
FE	30.4			10.7			3.3		
NREM (*Adpt.*)	G=5	G=10	G=15	G=5	G=10	G=15	G=5	G=10	G=15
F = 160	25.2	24.1	23.5	11.0	10.8	10.5	6.2	5.8	6.1
F = 320	23.2	21.2	21.0	9.8	9.5	9.5	5.9	5.9	5.6
F = 480	21.6	20.3	20.0	10.1	9.4	9.8	5.8	5.6	5.5
F = 640	20.2	18.5	18.4	**9.1**	9.2	9.4	5.8	5.6	5.6
F = 800	19.9	17.9	18.0	9.5	8.7	9.3	5.8	5.6	5.6
F = 1200	**19.0**	**17.5**	**17.2**	9.3	**8.4**	**8.9**	**5.6**	**5.5**	**5.3**

(b) Test set B

SNR(dB)	**-5**			**0**			**5**		
NREM (*Fixed*)	57.5			23.8			8.8		
GMM/HMM	64.0			25.9			7.4		
SC	52.4			23.5			11.0		
FE	52.6			20.5			5.7		
NREM (*Adpt.*)	G=5	G=10	G=15	G=5	G=10	G=15	G=5	G=10	G=15
F = 160	57.1	55.8	56.1	23.5	**23.1**	23.4	8.2	**8.0**	**8.2**
F = 320	55.6	56.2	55.9	23.4	**23.1**	23.5	8.2	8.4	8.8
F = 480	55.8	56.2	**55.7**	22.8	23.4	23.1	8.6	8.3	8.4
F = 640	**55.2**	56.4	**55.7**	22.8	**23.1**	23.0	8.2	8.3	8.7
F = 800	56.0	**55.7**	55.8	22.8	23.3	**22.7**	**7.9**	8.3	8.6
F = 1200	55.4	56.1	56.6	**22.1**	23.9	23.2	8.4	8.6	8.7

using *adaptive* and *fixed* noise dictionaries. The experiments with adaptive dictionaries are performed by varying F between 160 to 1200 and G between 5 to 15 exemplars per sequence. The results are given at the lower panel of Table 6.1a and 6.1b. In Table 6.1a, the recognition results obtained on test set A are shown. The baseline system using fixed dictionaries provides WERs of 47.1%, 21.2% and 9.3% at SNR level of -5, 0 and 5 dB respectively. The proposed adaptive noise modeling scheme with F=160 and G=5 training sequences reduces the WERs dramatically to 25.2%, 10.9% and 6.2% at the same SNR levels. For

G=10, the WER reduces to 24.1% at -5 dB. G=10 is a reasonable choice as increasing G further brings no significant improvement. At SNR levels of 0 and 5 dB, G has a less noticeable impact on the recognition accuracy. Increasing F provides further improvements on the recognition accuracy with WERs of 20.3%, 17.9% and 17.5% for F equal to 480, 800 and 1200 at SNR of -5 dB. The recognition results follow a similar trend at SNRs of 0 and 5 dB. The lower panel of Table 6.1b presents the recognition results for test set B. The baseline system using fixed dictionaries provides WERs of 57.5%, 23.8% and 8.8% at SNR level of -5, 0 and 5 dB respectively. For the mismatched noise case, the selection technique still provides some improvement for any G and F which is explained by the increased spectral diversity of the available noise exemplars. Unlike the matched case, increasing G or F does not have a considerable impact on the recognition accuracy.

The recognition accuracies provided by the baseline and the proposed systems on the development and test set of CHIME-2 data are presented in Table 6.2a and Table 6.2b. The results on development and test sets follow a similar pattern, thus, we focus only on the test set results. The baseline system using 400 sniffed exemplars provides 69.3%, 76.8% and 84.5% at SNRs of -6, -3 and 0 dB. The recognition system using only adaptive dictionaries with X=3 provides comparable results with 69.8%, 76.5% and 83.9% at the same SNR levels. The mixed dictionaries obtained by combining 200 sniffed exemplars (SE) with adaptive noise dictionaries having X=2 provide the best performance. This system provides 71.2%, 78.9% and 85.3% at SNRs of -6, -3 and 0 dB with an absolute improvement of 1.9%, 2.1% and 0.8% respectively. Another setup that gives promising results is the one using noise dictionaries with 300 SE and X=1. All setups using adaptive noise modeling provide comparable results at higher SNRs. The recognition results with higher X values are not reported as increasing X does not improve the results with an increased computational burden.

From these results, it can be concluded that the preliminary noise sequence selection technique benefits from the larger noise repository with a rather coarse sampling of the noise-only sequences in the repository. For AURORA-2 data, setting G=10 exemplars per sequence captures the within noise-only sequence variation well enough and larger G values do not improve the recognition accuracy. Finally, depending on the available memory, the noise repository size F can be increased further to have better coverage of the variation in background noise and hence improved performance. The experiments on CHIME-2 data show that combining sniffed exemplars with the exemplars extracted from the selected sequences provides superior noise modeling compared to only sniffing similar amounts of noise exemplars. Furthermore, it has been shown that the best recognition performance at lower SNR levels is achieved using 350-450

mixed noise exemplars per dictionary. Increasing the amount of noise exemplars further does not bring any improvement. This upper bound on the recognition performance is explained by the poor speech modeling provided by the speech dictionaries due to the limited amount of training data.

6.6 Conclusion

This chapter investigates the impact of several parameters of an exemplar-based adaptive noise modeling technique on the recognition accuracy. A non-negative sparse coding-based noise exemplar selection technique is described in the previous work that selects noise exemplars on-the-fly to be able to model the spectrotemporal content of the actual noise conditions. Using the optimal parameters, the final system with adaptive noise modeling uses less noise exemplars compared to the system using fixed dictionaries and provides better recognition accuracy on the AURORA-2 data. Moreover, the experiments on CHIME-2 data show that the mixed dictionaries containing sniffed and adaptively selected noise exemplars outperform the baseline using sniffed exemplars only. Overall, the proposed approach appears to be an effective noise dictionary design scheme that can be incorporated in exemplar-based ASR approaches.

Table 6.2: Recognition accuracies in percentages obtained on development and test set the CHIME-2 data - SE: Sniffed Exemplars

(a) Development Set

SNR(dB)	**-6**	**-3**	**0**	**3**	**6**	**9**
NREM (400SE)	69.4	76.4	85.0	90.1	92.9	93.3
GMM/HMM	49.3	58.6	67.5	75.0	78.8	82.9
SC	75.5	81.4	87.5	89.9	92.4	92.3
FE	68.0	72.2	80.9	86.7	89.0	90.5
NREM (*Adpt.*)	**-6**	**-3**	**0**	**3**	**6**	**9**
X = 1	64.8	71.3	80.6	86.8	90.9	92.4
X = 2	66.0	73.6	82.0	89.1	92.1	93.2
X = 3	69.1	76.3	84.6	89.3	92.3	93.5
X = 1 + 100SE	68.3	76.4	83.6	90.2	92.3	92.9
X = 2 + 100SE	67.0	73.9	83.0	**90.7**	92.3	93.4
X = 3 + 100SE	67.1	74.6	83.7	90.4	92.5	93.3
X = 1 + 200SE	65.7	73.9	83.6	90.3	92.4	93.3
X = 2 + 200SE	70.6	**78.0**	84.7	90.4	92.6	**93.8**
X = 1 + 300SE	**71.3**	77.8	**85.1**	90.3	**92.8**	93.6

(b) Test set

SNR(dB)	**-6**	**-3**	**0**	**3**	**6**	**9**
NREM (400SE)	69.3	76.8	84.5	88.8	91.9	93.5
GMM/HMM	49.7	57.9	67.8	73.7	80.8	82.7
SC	76.5	81.3	88.9	90.5	92.7	93.2
FE	67.2	75.9	81.1	86.4	90.7	92.0
NREM (*Adpt.*)	**-6**	**-3**	**0**	**3**	**6**	**9**
X = 1	65.5	72.2	80.8	86.4	89.8	93.1
X = 2	68.4	75.3	83.6	87.8	90.3	92.8
X = 3	69.8	76.5	83.9	87.8	90.5	92.7
X = 1 + 100SE	69.5	74.9	**85.3**	88.7	91.9	93.3
X = 2 + 100SE	68.0	75.3	84.5	87.7	91.8	92.7
X = 3 + 100SE	67.7	75.3	84.0	87.5	91.0	92.6
X = 1 + 200SE	67.2	74.9	**85.3**	87.0	92.4	93.2
X = 2 + 200SE	**71.2**	**78.9**	**85.3**	88.7	91.9	92.8
X = 1 + 300SE	70.6	77.4	**85.3**	**88.8**	**92.6**	**93.4**

Chapter 7

Alpha-Beta Divergence for N-REM

The noise robust exemplar matching (N-REM) framework performs automatic speech recognition using exemplars, which are the labeled spectrographic representations of speech segments extracted from training data. By incorporating a sparse representations formulation, this technique remedies the inherent noise modeling problem of conventional exemplar matching-based automatic speech recognition systems. In this framework, noisy speech segments are approximated as a sparse linear combination of the exemplars of multiple lengths, each associated with a single speech unit such as words, half-words or phones. On account of the reconstruction error-based back end, the recognition accuracy highly depends on the congruence of the speech features and the divergence metric used to compare the speech segments with exemplars. In this work, we replace the conventional Kullback-Leibler divergence (KLD) with a generalized divergence family called the Alpha-Beta divergence with two parameters, α and β, in conjunction with mel-scaled magnitude spectral features. The proposed recognizer traverses the (α,β) plane depending on the amount of contamination to provide better separation of speech and noise sources. Moreover, we apply our recently proposed active noise exemplar selection (ANES) technique in a more realistic scenario where the target utterances are degraded by genuine room noise. Recognition experiments on the small vocabulary track of the 2^{nd} CHiME Challenge and the AURORA-2 database have shown that the novel recognizer with the AB divergence and ANES outperforms the baseline system using the generalized KLD with tuned sparsity, especially at lower SNR levels.

This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo

Van hamme, "*Noise Robust Exemplar Matching with Alpha-Beta Divergence*", Submitted to Speech Communication, 2015.

7.1 Introduction

Data-driven automatic speech recognition (ASR) techniques [1, 29, 36, 70, 143, 159, 161] became popular in the last decade as a viable alternative after the long dominance of statistical acoustic modeling in the form of the Gaussian mixture models (GMM) in hidden Markov models (HMM) [15]. Templates or exemplars are labeled speech segments of multiple lengths extracted from training data, each associated with a certain speech unit such as phones, syllables or words. As they preserve the complete duration and trajectory information, exemplars are more immune to the inherent spectrotemporal variation of speech and its deteriorating effect on the ASR [13] compared to the conventional GMM/HMM- or deep neural networks (DNN)-based recognition systems. Moreover, it has been shown that using reasonably large exemplar sets overcomes the well-known generalization problem of the previous exemplar-based approaches [30, 149, 158].

Exemplar matching-based recognition can be performed by evaluating the similarity of the exemplars with the segments from the input speech with respect to a distance/divergence metric by applying dynamic time warping [30, 129, 144]. In these applications, speech is represented using discriminatively trained features to ensure that the used distance/divergence metric mostly yields lower scores for the matching class compared to the other classes, resulting in increased recognition accuracies. The input speech segments can be simply classified as the label of the closest exemplar, or by a voting scheme on the set of K nearest neighbors [57].

Exemplar-based sparse representations (SR) is an alternative data-driven ASR approach in which the spectrogram of input speech segments is modeled as a sparse linear combination of exemplars. SR-based techniques have been successfully used for speech enhancement [54], feature extraction [142] and speech recognition [50, 92, 162]. These approaches model the acoustics using same-length exemplars labeled on the frame level and stored in a single overcomplete dictionary. The exemplar weights are obtained by solving a regularized convex optimization problem with a cost function consisting of the approximation quality with respect to a divergence and a term to induce sparse linear combinations using only a few exemplars. The choice of the divergence depends on the used speech features (how speech and noise sources are distributed in the high-dimensional feature space) to obtain reasonable sparse linear combinations. The non-negativity requirement of the SR formulation

prevents the use of discriminatively trained features in this framework. The generalized Kullback-Leibler divergence (KLD) with the mel scaled magnitude spectral features has been successfully used in various applications in source separation, SR-based noise robust speech recognition and polyphonic music transcription [138, 151, 152, 162, 174]. King et al. investigated the optimal parameter of the beta divergence as a cost function for non-negative matrix factorization-based speech separation and music interpolation in [97].

This chapter focuses on the divergence used by a recently proposed exemplar matching-based recognition approach, dubbed noise robust exemplar matching (N-REM) [184], which performs conventional exemplar matching in a SR formulation to be able to model noisy speech. Similar to the exemplar matching approaches, N-REM uses exemplars associated with a single speech unit such as phones, syllables, half-words or words. These exemplars are organized in separate dictionaries based on their duration (frame length) and class (associated speech unit). By applying a sliding window approach, the noisy speech segments are jointly approximated as a linear combination of the speech and noise exemplars using each dictionary. The recognizer adopts a reconstruction error based back-end, i.e. the recognition is performed by comparing the approximation quality for different classes quantified by a divergence measure and choosing the class sequence that minimizes the total reconstruction error.

The divergence plays an essential role in the recognition performance of N-REM on account of the reconstruction error based backend. The optimal divergence is expected to weight the individual reconstruction errors of each time-frequency cells in a way that the most informative cells contribute the most to the total reconstruction error. In this work, we use the Alpha-Beta (AB) divergence [20] in place of the generalized KLD to quantify the approximation error. The AB divergence is a family of divergences with two parameters, namely α and β. For different values of these parameters, the AB divergence connects various well-known distance/divergence measures such as the Euclidean distance, Hellinger distance, Itakura-Saito divergence and generalized KLD. The higher degree of freedom offered by the AB divergence has been shown to enable better robustness against noise and outliers [20]. The initial ASR results at lower SNR levels are presented in [185] and it has been shown that using AB divergence with an appropriate (α,β) pair provides better recognition than the generalized KLD with tuned sparsity.

The main contribution of this chapter is a novel noise robust recognizer which traverses the (α,β) plane based on the estimated SNR level to perform the most accurate separation of speech and noise sources. The recognition performance of the proposed recognizer is investigated on the small vocabulary track of the 2$^{\text{nd}}$ CHiME Challenge (CHIME-2) and the AURORA-2 database. Secondly, an in-depth discussion on the impact of the divergence parameters on the recognition

performance is provided by comparing the behaviour of the generalized KLD and AB divergence for several (α,β) pairs. Finally, we apply the adaptive noise modeling technique, active noise exemplar selection (ANES) [184], on the CHIME-2 data to investigate the recognition performance in case of genuine room noise. The rest of the chapter is organized as follows. The N-REM using the AB divergence is described in Section 7.2. Section 7.3 discusses the evaluation setup and implementation details. Section 7.4 presents the recognition results and a discussion about the results is given in Section 7.5. Section 7.6 provides a general discussion and the concluding remarks.

7.2 Noise Robust Exemplar Matching

N-REM models noisy speech segments as a sparse linear combination of speech and noise exemplars of various lengths that are stored in multiple dictionaries. The overview of the recognizer is given in Figure 7.1. Compared to a system using fixed-length exemplars stored in a single dictionary, using separate dictionaries for each class results in noisy speech segments being approximated as a linear combination of exemplars belonging to the same class only. From the geometrical interpretation of SR-based source separation, it is known that the farther the convex hull of the basis vectors belonging to different sources (speech and noise in this case) are, the better the separation is [37]. Hence, the use of separate dictionaries for each speech unit provides a more precise representation in the high-dimensional feature space.

7.2.1 Model Description

Training frame sequences representing various speech units (speech exemplars) are extracted based on the state-level alignments obtained using a conventional HMM-based recognizer. Speech exemplars, each comprised of D mel frequency bands and spanning l frames, are reshaped into a single vector and stored in the columns of a speech dictionary $\mathbf{S}_{c,l}$: one for each class c and each frame length l. Each dictionary is of dimensionality $Dl \times N_{c,l}$ where $N_{c,l}$ is the number of available speech exemplars of class c and frame length l. Similarly, a noise dictionary $\mathbf{N}_l$ for each frame length l is formed by reshaping the noise exemplars. Each speech dictionary is concatenated with the noise dictionary of the same length to form a combined dictionary $\mathbf{A}_{c,l} = [\mathbf{S}_{c,l}\, \mathbf{N}_l]$ of dimensionality $Dl \times M_{c,l}$ where $M_{c,l}$ is the total number of available speech and noise exemplars.

An observed noisy (and/or reverberated) speech segment of frame length T frames is also reshaped into vectors by applying a sliding window approach [50]

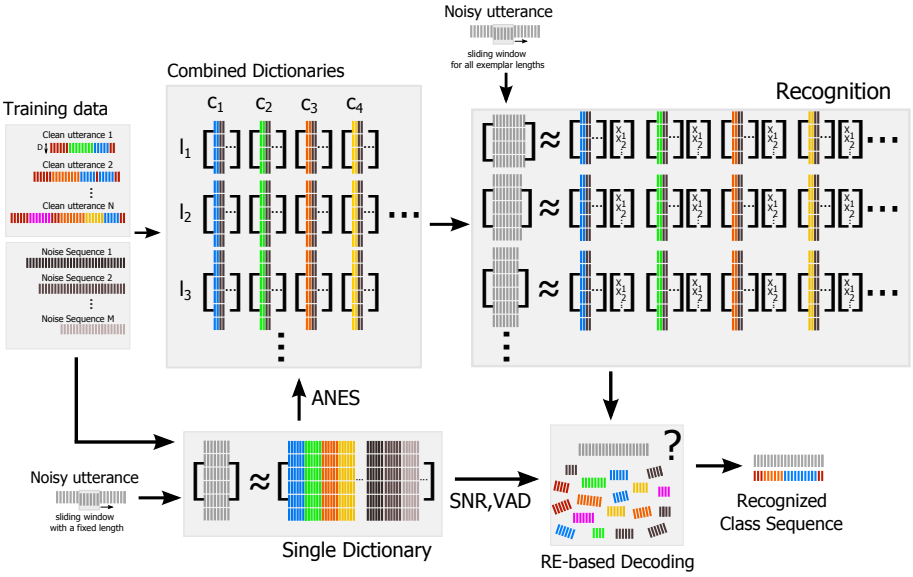


Figure 7.1: The Recognizer Overview. The single dictionary is used for the VAD, SNR estimation and active noise exemplar selection (ANES). Noise exemplars that are used in the recognition are selected based on the single dictionary. Speech exemplars are extracted from the training data using the segmentation information. They are organized in dictionaries based on their length and class. Noise dictionaries are concatenated to the speech dictionaries forming the combined dictionaries. Non-negative sparse coding (NSC) is applied to approximate noisy test utterances using the combined dictionaries. After a fixed number of iterations, the reconstruction errors are calculated and a dynamic programming algorithm is applied to find the class sequence with the minimum reconstruction error.

with window length of l frames and stored in an observation matrix $\mathbf{Y}_l = [\mathbf{y}_l^1, \mathbf{y}_l^2 ..., \mathbf{y}_l^{(T-l+1)}]$ of dimensionality $Dl \times (T-l+1)$. Due to multiple-length exemplars, the window length l is varied between the minimum exemplar length $l_{\min}$ and maximum exemplar length $l_{\max}$ yielding observation matrices $\mathbf{Y}_l$ for $l_{\min} \leq l \leq l_{\max}$. For every class c, each observation vector $\mathbf{y}_l$ is expressed as a linear combination of the exemplars that are stored in the dictionaries of the same length:

$$\mathbf{y}_l \approx \sum_{m=1}^{M_{c,l}} x_{c,l}^m \mathbf{a}_{c,l}^m = \mathbf{A}_{c,l}\mathbf{x}_{c,l} \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \tag{7.1}$$

where $\mathbf{x}_{c,l}$ is an $M_{c,l}$-dimensional non-negative weight vector. The sparse

solutions of $\mathbf{x}_{c,l}$ yield a more realistic approximation of the observed segments without overfitting and have been shown to provide better recognition results [80, 174]. The combined dictionaries consisting of speech and noise exemplars are presumed to model all acoustic variability in the observed signal due to pronunciation variation, background noise and so forth. This model can also model reverberation by storing reverberated speech exemplars rather than clean speech exemplars.

7.2.2 Finding Exemplar Weights

The exemplar weights $\mathbf{x}_{c,l}$ are obtained by minimizing the cost function consisting of a single term which quantifies the approximation error $d(\mathbf{y}_l, \mathbf{A}_{c,l}\mathbf{x}_{c,l})$ for non-negative exemplar weights. Unlike the baseline recognizer and the other SR-based approaches, the new cost function does not enforce sparsity on the exemplar weights. The impact of the missing sparsity inducing term is investigated in Section 7.4.1 by visualizing the sparseness of the obtained exemplar weights that are obtained by only minimizing the approximation error. This optimization problem can be solved with the non-negative sparse coding (NSC) [79].

The value of the approximation error is highly dependent on the divergence measure d and the representation of speech and noise sources. Particularly, the adopted divergence measure is expected to provide more reliable reconstruction errors by emphasizing the reliable and informative time-frequency bins which are dominated by the desired source (speech in this case). Prior work has shown that the mel-scaled spectral features provide better source separation when used in conjunction with the generalized KLD compared to the Euclidean distance [174].

Recently, the AB divergence has been proposed and its application as a cost function for non-negative matrix factorization has been investigated [20]. Motivated by its capabilities to weight and scale the individual ratios of the noisy speech and its approximation, $\mathbf{y}_l^i / \hat{\mathbf{y}}_{c,l}^i$ where $\hat{\mathbf{y}}_{c,l} = \mathbf{A}_{c,l}\mathbf{x}_{c,l}$, we investigate the recognition performance of the proposed system using the AB divergence for d. The influence of different (α,β) values on this ratio is detailed in [20].

The AB divergence is defined as

$$d_{AB}^{(\alpha,\beta)}(\mathbf{y},\hat{\mathbf{y}}) = \begin{cases} -\frac{1}{\alpha\beta}\sum_{k=1}^{K}\left(y_k^{\alpha}\hat{y}_k^{\beta} - \frac{\alpha}{\gamma}y_k^{\gamma} - \frac{\beta}{\gamma}\hat{y}_k^{\gamma}\right) & \text{for } \alpha,\beta,\gamma \neq 0, \\ \frac{1}{\alpha^2}\sum_{k=1}^{K}\left(y_k^{\alpha}\log(\frac{y_k^{\alpha}}{\hat{y}_k^{\alpha}}) - y_k^{\alpha} + \hat{y}_k^{\alpha}\right) & \text{for } \alpha \neq 0, \beta = 0 \\ \frac{1}{\alpha^2}\sum_{k=1}^{K}\left(\log(\frac{\hat{y}_k^{\alpha}}{y_k^{\alpha}}) + \frac{y_k^{\alpha}}{\hat{y}_k^{\alpha}} - 1\right) & \text{for } \alpha = -\beta \neq 0, \\ \frac{1}{\beta^2}\sum_{k=1}^{K}\left(\hat{y}_k^{\beta}\log(\frac{\hat{y}_k^{\beta}}{y_k^{\beta}}) - \hat{y}_k^{\beta} + y_k^{\beta}\right) & \text{for } \alpha = 0, \beta \neq 0, \\ \frac{1}{2}\sum_{k=1}^{K}\left(\log(y_k) - \log(\hat{y}_k)\right)^2 & \text{for } \alpha,\beta = 0 \end{cases} \tag{7.2}$$

where $\gamma = \alpha+\beta$. The two parameters of the AB divergence can be automatically adjusted based on the amount of contamination in the target utterance as the recognition performance for different noise levels depends on the emphasized (reliable) time-frequency bins. For the NSC solution, we apply the relaxation and non-linear projection techniques proposed in [21] for faster convergence to the multiplicative update rule derived in [20] to minimize the approximation error. The multiplicative update rule which minimizes the approximation error using the AB divergence for $\alpha \neq 0$ is given by

$$\mathbf{x}_{c,l} \leftarrow (\mathbf{x}_{c,l} \odot ((\mathbf{A}_{c,l}^T\mathbf{Z}_{c,l}) \oslash (\mathbf{A}_{c,l}^T(\mathbf{A}_{c,l}\mathbf{x}_{c,l})^{\cdot[\gamma-1]}))^{\cdot[\omega/\alpha]})^{\cdot[1+\theta]}, \tag{7.3}$$

where $\mathbf{Z}_{c,l} = \mathbf{y}_l^{\cdot[\alpha]} \odot (\mathbf{A}_{c,l}\mathbf{x}_{c,l})^{\cdot[\beta-1]}$ and $^{\cdot[\ ]}$ denotes element-wise exponentiation. ω is a value between $(0, 2)$ and θ is a very small positive number [21]. $\mathbf{1}$ is a Dl-dimensional vector with all elements equal to unity.

7.2.3 Decoding

All observation matrices Y_l for $l_{\min} \leq l \leq l_{\max}$ are approximated using the combined dictionaries $\mathbf{A}_{c,l}$ of the corresponding length by applying the multiplicative update rule in Equation (7.3). To quantify the approximation quality, we use the reconstruction error between the noisy speech segments and their approximations. The multiplicative update rule is applied iteratively until the reconstruction error provides enough discrimination between different classes. The number of iterations that satisfies this criterion has been investigated in pilot experiments. After a fixed number of iterations for all dictionaries, the reconstruction errors between the observation matrix $\mathbf{Y}_l$ and its approximations $\mathbf{A}_{c,l}\mathbf{x}_{c,l}$ are calculated for $l_{\min} \leq l \leq l_{\max}$. As the label of each dictionary is known, decoding is performed by applying a single-stage dynamic programming algorithm [128] to find the class sequence that minimizes the reconstruction

error (taking the grammar into account if necessary). This search problem is visualized as a three-dimensional grid search over grid points (x, y, z) which are defined by the time frames x of a noisy speech segment, time frames y of its approximation and the dictionary number z [128]. Noisy speech segments are only matched with the dictionaries of the same duration, i.e. no time warping is performed.

7.2.4 Preprocessing of Noisy Speech

Before the recognition phase, the noisy speech is approximated using a single overcomplete dictionary to gather some information about the target utterance such as the voice activity detection (VAD), signal-to-noise ratio (SNR) estimation and noise characteristics. This single dictionary is formed by choosing an exemplar length L_s containing a vast number of samples from each class. The single speech dictionary $\mathbf{S}^*_{L_s}$ contains speech exemplars of all classes with the same length. The single noise dictionary $\mathbf{N}^*_{L_s}$ has noise exemplars that are extracted from the noise-only training sequences. The preprocessing step performs non-negative sparse coding using the single (combined) dictionary $\mathbf{A}^*_{L_s} = [\mathbf{S}^*_{L_s} \mathbf{N}^*_{L_s}]$

$$\mathbf{Y}_{L_s} \approx \mathbf{A}^*_{L_s} \mathbf{x}_{L_s} \qquad \text{s.t.} \qquad \mathbf{x}_{L_s} \geq 0. \tag{7.4}$$

where $\mathbf{Y}_{L_s}$ is the observation matrix having a window length of L_s frames. As the proposed recognizer uses an SNR-dependent (α, β) pair, the generalized KLD is used as a reference for obtaining the weights of exemplars in the single dictionary. The multiplicative update rule for finding the exemplar weights $\mathbf{x}_{L_s}$ can be found in [184].

The information provided by the exemplar weights $\mathbf{x}_{L_s}$ are used for multiple purposes. Firstly, a known problem of SR approaches working on magnitude spectra is that the silence exemplars are hard to recognize: perfect silence is modeled with zero weights of all exemplars [50]. In a practical noisy mixture, it is well-approximated by combining speech and noise exemplars with small weights, thus all classes will score equally well. To overcome this problem, the reconstruction errors belonging to the silence dictionaries have to be compensated for the noisy speech segments which do not contain speech. For this purpose, the recognizer embodies the preprocessing step to perform VAD for predicting whether a noisy speech segment contains speech and to estimate the SNR level for adjusting the amount of compensation. An indicator of the SNR level, SNR_{ind}, is calculated as the ratio of total speech weights and total

speech and noise weights is used in order to limit the range to $[0, 1]$,

$$\mathrm{SNR}_{\mathrm{ind}} = \frac{\sum_{w=1}^{W} \sum_{m=1}^{J} x_{L_s}^{w,m}}{\sum_{w=1}^{W} \sum_{m=1}^{M} x_{L_s}^{w,m}}. \tag{7.5}$$

$\mathbf{x}_{L_s}^{w}$ is the sparse weight vector corresponding to w^{th} of W noisy segments of length L_s. J is the number of speech exemplars and M is number of all exemplars. The details of the silence compensation can be found in [184].

The preprocessing step also provides useful information about the spectrotemporal content and the level of the background noise. The former information is used for extracting a small set of noise exemplars that are able to model the actual noise conditions by applying the adaptive noise exemplar selection (ANES) technique. The level of the background noise is quantified by the estimated SNR level and the number of noise exemplars that are used in the recognition phase is chosen based on this estimated SNR value. In practice, the recognizer uses more noise exemplars for lower SNR levels and less or no exemplars for higher SNR levels. This way of noise modeling has been shown to both reduce the computational complexity and improve the recognition accuracies at higher SNRs. The adaptive and SNR-dependent noise modeling approach is detailed in [184] and summarized in Section 7.2.5.

Finally, the proposed recognizer chooses the divergence parameters (α, β) according to the estimated SNR value to provide better separation and improve the recognition performance. The path providing the best recognition performance on the (α, β) plane is determined in advance on development data and the divergence parameters are chosen on this predetermined path based on the estimated SNR level.

7.2.5 Speech and Noise Dictionaries

Several dictionary design techniques have been applied for effective speech and noise modeling using the exemplars. As the speech exemplars are associated with a single speech unit, their length distribution is class-dependent which results in unevenly populated speech dictionaries. Speech dictionary design mainly involves increasing the number of exemplars in underpopulated speech dictionaries to avoid poor acoustic modeling. *Prewarping* [183] is applied to increase the number of the exemplars by removing a small number of frames, excluding the very first and last frame, from an exemplar of length l to obtain shorter exemplars of length $l_{\mathrm{new}} < l$.

Noise exemplars are extracted from noise-only training sequences for arbitrary length. While there are a vast number of noise exemplars for every exemplar length, only the ones that match the actual noise conditions will be useful during the recognition. As a result, noise dictionary design mainly focuses on accurate modeling of the background noise using the smallest possible number of noise exemplars. ANES is an adaptive way of noise modeling that accurately picks a small number of noise exemplars that can model the actual noise conditions. Large performance gains have been reported compared to fixed noise modeling, i.e. using the same set of noise exemplars for all test utterances, especially at lower SNRs [184]. Adaptive noise dictionaries are obtained based on the noise weights that are provided by the single dictionary setup described in Section 7.2.4. This technique aims to select a small number of noise exemplars that can accurately model the actual noise conditions. An equal number of exemplars is extracted from a large number of noise-only training sequences and stacked in a single noise dictionary. In order to identify which noise-only training sequences can accurately model the actual noise conditions, all weights belonging to the noise exemplars extracted from each noise-only training sequence are accumulated. With the same motivation as discussed in [182], noise exemplars used for the recognition are extracted from the most active noise-only training sequences, i.e. the sequences with the highest weights. For the details of the ANES technique, we refer the reader to [184].

Another technique that has been proposed for improved noise modeling in SR-based recognition systems is called *noise sniffing* [53]. This technique acquires noise exemplars on the fly from the immediate neighborhood of the target utterance. The extracted noise exemplars are added to the combined dictionaries and used for the recognition. In case of limited noise context, a small number of frames from the beginning and end of the target utterance are extracted and contained in combined dictionaries. Shifted copies of these frame sequences are also included to provide some degree of shift-invariance [52]. The VAD information is used to detect the speech onset and offset points.

SNR-dependent noise modeling approach finds a compromise between the accuracy of the noise modeling and computational complexity by adjusting the amount of the noise exemplars in the combined dictionaries depending on the estimated SNR level. At lower SNRs, a larger number of noise-only training sequences are used for noise exemplar extraction. Consequently, computational complexity of the recognizer is reduced at high SNRs without loss of recognition accuracy while preserving the noise modeling capabilities at lower SNRs. Moreover, SNR-dependent noise modeling provides gains in the recognition accuracy of clean speech, as the dictionaries contain only a few noise exemplars during the recognition of clean speech.

7.3 Experimental Setup

7.3.1 Databases

AURORA-2

The recognition performance of N-REM is further evaluated on the test set A and B of the AURORA-2 corpus [77]. The training material of AURORA-2 consists of a clean and a multi-condition training set, each containing 8440 utterances with one to seven digits in American English. The multi-condition training set was constructed by mixing the clean utterances with noise at SNR levels of 20, 15, 10 and 5 dB.

Test set A consists of 4 clean and 24 noisy datasets with four noise types (subway, babble, car and exhibition) at six SNR levels, 20, 15, 10, 5, 0 and -5 dB. The noise types of this test set match the multi-condition training set. Test set B has the same number of test sets with four different noise types (restaurant, street, airport, station) at the same SNR levels. Each subset contains 1001 utterances. To reduce the simulation times, we subsampled the test sets by a factor of 4 (250 utterances per test set, 1000 utterances per SNR). A different subset with 100 utterances from each test set is used for development purposes. All data has a sampling frequency of 8 kHz.

CHIME-2

The small vocabulary track of the 2$^{\text{nd}}$ CHiME Challenge [172] addresses the problem of recognizing commands in a noisy living room. The clean utterances in the CHIME-2 data are taken from the GRID corpus [23] which contains utterances from 34 speakers reading 6-word sequences of the form *command-color-preposition-letter-digit-adverb*. There are 25 different letters, 10 different digits and 4 different alternatives for each of the other classes. Even though there is no silence between the words, leading silences of variable duration exist occasionally. The recognition accuracy of a system is calculated based on the correctly recognized letter and digit keywords.

The clean utterances are convolved with binaural room impulse responses with speaker head movement effects which are recorded in a living room. Then, the resulting reverberated utterances are mixed with binaural recordings of genuine room noise recorded in the same living room at SNR levels of 9, 6, 3, 0 ,-3 and -6 dB. The training set contains 500 utterances per speaker (17,000 utterances in total) with clean, reverberated and noisy versions. Noisy utterances are provided

both in isolated or embedded form. Embedded recordings contain 5 seconds of background noise before and after the target utterance. The development and test sets contain 600 utterances from all speakers at each SNR level (3600 utterances in total for each set) both in isolated and embedded from. The immediate noise context of the target utterances are available in 164 embedded recordings in the development set and 176 embedded recordings in the test set. All data has a sampling frequency of 16 kHz.

7.3.2 Dictionary Creation and Implementation Details

AURORA-2

The speech exemplars are extracted from the clean training set. Acoustic feature vectors are represented in mel-scaled magnitude spectra with 23 frequency bands. The speech exemplars representing half-digits are segmented by a conventional HMM-based system. The recognizer uses in total 52295 speech exemplars excluding 990 silence exemplars. The number of noise exemplars varies depending on the duration of the noise-only sequences that are selected by ANES. On average, the recognizer uses 11355 and 1044 noise exemplars/utterance in total at SNR level of -5 dB and clean speech respectively. The minimum and maximum exemplar lengths are 8 and 40 frames respectively. Exemplars longer than 40 frames are omitted to limit the number of dictionaries.

The single noise dictionary contains noise exemplars that are extracted from 800 longest noise-only training sequences (50 sequences from each multicondition training set). From each noise-only training sequence, 10 noise exemplars are extracted with equal frame shifts resulting in 8000 noise exemplars. The single speech dictionary contains 2200 speech exemplars (100 exemplars from each class excluding silence). The speech and noise exemplars contain 15 frames. The first and last 20 frames of the target utterances are assumed not to contain speech and 150 noise exemplars with 15 frames (5 exemplars and 70 shifted copies from each end) are extracted as described in [184] and concatenated to the single dictionary. The speech and noise exemplar weights are obtained after 300 iterations.

In the recognition phase, noise dictionaries are created by performing noise sniffing and active noise exemplar selection. The details of the noise dictionary creation are given in [184]. The recognizer uses in total 675 dictionaries of 23 different classes (half-digits plus silence). The combined dictionaries and observation matrices are l_2-normalized for all SNR levels. The multiplicative update rule is iterated 100 times for convergence of all frame lengths with $\omega = 1.75$ and $\theta = 0.008$. The divergence parameters (α, β) providing the best

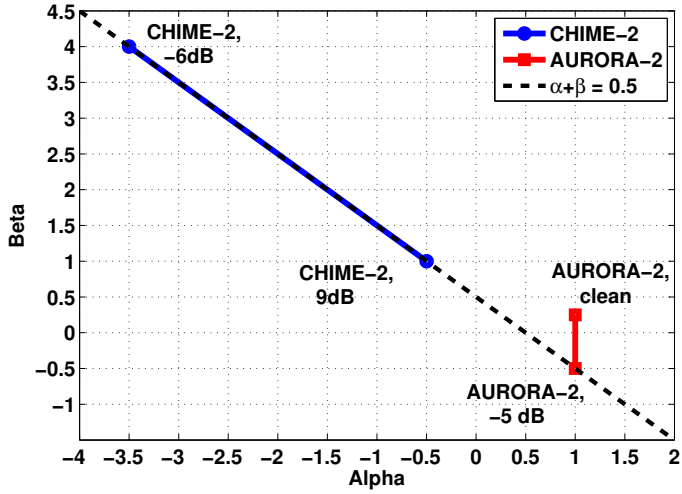


Figure 7.2: The line segments on the AB plane used for the recognition of the AURORA-2 and CHIME-2 databases - The α+β=0.5 line is also visualized which provided the best results for noisy conditions on both databases

performance at the lowest and highest SNR are investigated on the development data. The pilot experiments on the development data have shown that the best results are obtained on the line $\alpha = 1$ for the AURORA-2 database. The AB divergence with $(1, -0.5)$ and $(1, 0.25)$ provided the best recognition accuracies at SNR level of -5 dB and on clean speech. The line segments used during the recognition of both databases are illustrated in Figure 7.2 on the AB plane. A suboptimal estimation of the β value is performed in the interval of $[-0.5, 0.25]$ as a linear function of the SNR_{ind} value,

$$\beta = \max(\min(2 \cdot \text{SNR}_{\text{ind}} - 0.55), 0.25), -0.5). \tag{7.6}$$

CHIME-2

The exemplars and noisy speech segments are represented as mel-scaled magnitude spectral features extracted with a 26 channel mel-scaled filter bank ($D = 26$). The frame length is 25 ms and the frame shift is 10 ms. The binaural data is averaged in the spectral domain to obtain 26-dimensional feature vectors. The exemplars are extracted from the reverberated utterances in the training set according to the state-based segmentations obtained using the acoustic models in the toolkit provided with the database. Exemplars belonging to

each speaker are organized in separate dictionary sets for speaker-dependent modeling yielding 34 different dictionary sets. Based on the availability of the exemplars, the minimum and maximum exemplar lengths are 4 and 40 frames respectively.

Half-word exemplars seemed to generalize sufficiently to unseen data. Half-word exemplars are extracted by cutting the word exemplars at the HMM state yielding the minimum average length difference between the two halves. Dictionary sizes vary with different classes and speakers. *Prewarping* is applied to boost the modeling capabilities of the underpopulated speech dictionaries (especially for the ones belonging to letters due to the high number of alternatives and hence the small number of exemplars per class) and it is limited to a single frame. The number of exemplars in each dictionary after prewarping is limited to 50. Further exemplar extraction details can be found in [184].

The noise dictionaries used for the recognition contain 200 noise exemplars that are acquired on the fly from the immediate neighborhood of the target utterance in both directions until the frames belonging to other target utterances. In addition to these sniffed noise exemplars, 200-300 noise exemplars are extracted from the most active 2 noise-only sequences selected by ANES. These noise exemplars are extracted with jumps of 3 frames yielding a different number of noise exemplars depending on the length of the noise-only sequence. Each embedded utterance is segmented into noise-only sequences by removing all target utterances. 5 noise exemplars of 25 frames are extracted from each noise-only sequence and stored in the single noise dictionary. The size of the single noise dictionary varies depending on the number of available noise-only sequences for each embedded recording. ANES only evaluates the noise-only sequences that are extracted from the embedded recording which contains the target utterance. The single speech dictionary contains 2354 full-word exemplars (maximum 50 exemplars from 51 classes) of 25 frames. The full-word exemplars are used in the single speech dictionary, as there is no exemplar length L_s containing a vast number of samples from each half-word class. The multiplicative update rule is iterated 25 times to obtain the exemplar weights. ω and θ are set to 1.75 and 0.008 respectively. The columns of the combined dictionaries and observation matrices are l_2-normalized. To investigate the impact of the divergence parameters, we have performed recognition experiments on the lowest and highest SNR levels of the development data. The best results at -6 dB and 9 dB are obtained using AB divergence with $(-3.5, 4)$ and $(-0.5, 1)$ respectively. Considering the results reported in [185], the divergence parameters are chosen on the line $\alpha + \beta = 0.5$ in the interval of $([-3.5, -0.5], [4, 1])$ as a linear function of the $\mathrm{SNR}_{\mathrm{ind}}$ value,

$$\alpha = \max(\min(7.5 \cdot \mathrm{SNR}_{\mathrm{ind}} - 5.75), -0.5), -3.5), \tag{7.7}$$

$$\beta = \max(\min(7.5 \cdot \mathrm{SNR}_{\mathrm{ind}} + 6.25), 4), 1). \tag{7.8}$$

7.3.3 Evaluation Metrics

We have opted for the metrics which have been traditionally used for the evaluation of the databases described in Section 7.3.1 for comparability with the previous literature. The word error rate has been used to quantify the recognition accuracy for the AURORA-2 digit recognition task. The keyword recognition accuracy is used to evaluate the system performance on the CHIME-2 data.

7.4 Results and Discussion

In this section, we firstly investigate the impact of the missing sparsity inducing term in the cost function by visualizing how sparse the exemplar weights for a few noisy utterances from test set A of the AURORA-2 database and report the recognition accuracies on the same dataset. Then, we compare the recognition performance of N-REM using the generalized KLD with induced sparsity with and without the adaptive noise modeling technique ANES on the CHIME-2 data. Finally, the recognition accuracies provided by N-REM using the AB divergence on the CHIME-2 data are presented. The recognition accuracies on both databases are compared with the baseline N-REM recognizer which uses the generalized KLD with tuned sparsity and some other comparable recognition schemes such as exemplar-based sparse representations approaches and a multicondition-trained HMM recognizer.

7.4.1 AURORA-2

Sparseness of AB Divergence

The induced sparsity has been a requirement for previous SR approaches using an overcomplete dictionary to select only a few exemplars with non-zero weights among thousands. Consequently, a realistic linear approximation of noisy speech segments are obtained without overfitting. On the other hand, N-REM uses dictionaries that contain a lot less exemplars than the ones used by the previous SR approaches and we investigate whether the inherent sparsity imposed due to the non-negativity constraint is enough for realistic approximations. In Figure 7.3, the largest exemplar weights obtained for 400 test utterances at -5 dB using

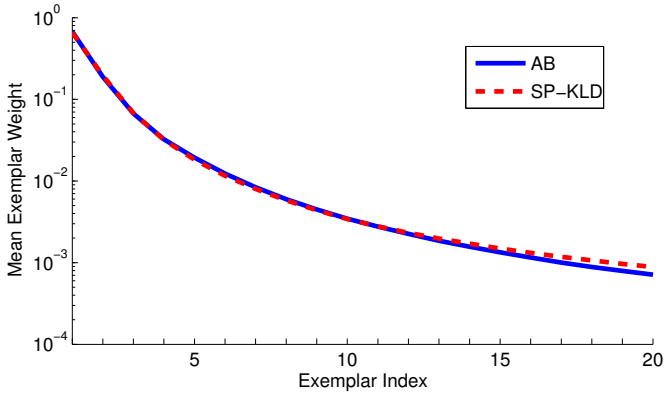


Figure 7.3: The comparison of exemplar weights obtained using the generalized KLD with tuned sparsity and the AB divergence - The weights are obtained using the 400 utterances used for development purposes at -5 dB in test set A of AURORA-2

the AB divergence are averaged and compared with the ones obtained using the generalized KLD with tuned sparsity on the same utterances. The sparseness of the exemplar weights is further visualized in Figure 7.4 by randomly picking two single-digit utterances at the same SNR level corrupted with subway and exhibition hall noise. For each noisy speech segment, the exemplars with the 5 largest weights are listed to observe how fast the weights are decaying for the dictionaries yielding the smallest reconstruction error. The visualized linear approximations have provided the minimum reconstruction error for each half-word. The divergence parameters are estimated based on the SNR_{ind} value. From this figure, it can be concluded that the linear combinations yielded by the multiplicative update rule given in Equation (7.3) are sparse enough to realistically estimate noisy segments due to the limited amount of exemplars in the dictionaries and the relaxation and non-linear projection techniques resulting in faster convergence.

Recognition Results

The recognition performance of N-REM using the AB divergence is compared with the baseline N-REM using the generalized KLD with tuned sparsity [184], a standard HMM recognizer [50] and two noise robust SR-based recognition techniques using fixed-length exemplars in a single overcomplete dictionary, namely sparse classification (SC) and feature enhancement (FE) [50]. The

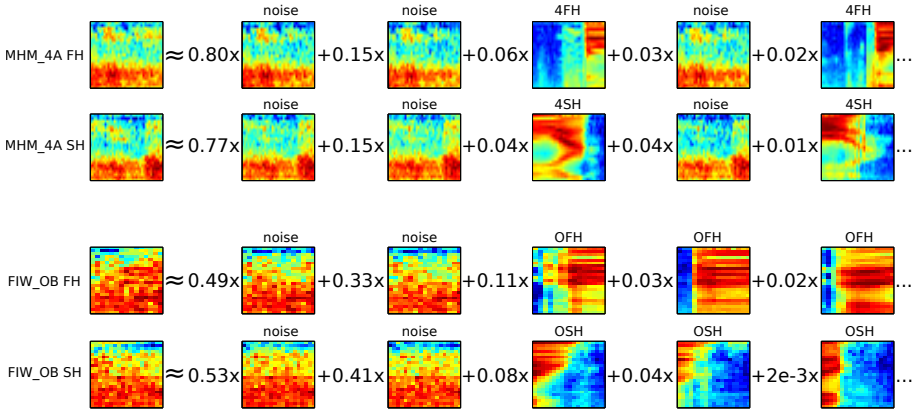


Figure 7.4: Illustration of the sparsity the exemplar weights provided by N-REM dictionaries using the AB divergence - The mel-scaled spectral patches given in the first column are the noisy mixtures extracted from noisy utterances MHM_4A and FIW_OB with subway and exhibition hall noise at an SNR level of -5 dB respectively. The following columns list the exemplars with the highest weights that are used to approximate the noisy segments in the first column. The label of each exemplar is given for each exemplar ('FH': first-half, 'SH': second-half)

SR-based recognition techniques achieve among the best known results on AURORA-2, especially at lower SNRs, performing significantly better than for instance the ETSI advanced front-end (AFE) which has been considered as a reference for the AURORA-2 database [78]. The HMM and FE recognition systems are trained on the multi-condition training set. The overcomplete dictionary used by SC and FE recognizers contain 10000 speech and 5000 noise exemplars with exemplar length of 30 frames. The exemplar weights are obtained after 600 iterations. We have performed recognition experiments on the same subset containing 1000 utterances from each SNR to obtain comparable recognition results.

The word error rates (WER) obtained on the test set A and B are given in Table 7.1. The upper panel presents the WER results provided by the baseline and proposed recognizers. The baseline N-REM provides WERs of 19.1% and 9.2% at SNR levels of -5 dB and 0 dB. The proposed system performs better than the baseline with WERs of 14.9% and 8.5% at the same SNR levels with an absolute improvement of 4.3% and 0.7%. These WERs are substantially lower than 35.2% and 13.8% of the SC recognizer and 30.4% and 10.7% of the FE recognizer.

Table 7.1: Word error rates in percentages obtained on test set A and B of the AURORA-2 database

(a) Test set A

SNR(dB)	clean	-5	0	5	10	15	20	*0-20*
N-REM (SP-KLD,ANES)	**1.7**	19.1	9.2	5.9	4.9	3.6	2.4	5.2
N-REM (AB,ANES)	1.8	**14.9**	**8.5**	**5.8**	**4.7**	**3.5**	**2.3**	**5.0**
HMM	0.7	60.8	24.3	7.3	2.9	1.3	0.8	7.3
SC	3.7	35.2	13.8	7.4	5.6	4.8	4.5	7.2
FE	0.5	30.4	10.7	3.3	1.5	1.1	0.7	3.5

(b) Test set B

SNR(dB)	clean	-5	0	5	10	15	20	*0-20*
N-REM (SP-KLD,ANES)	**1.7**	55.0	**24.3**	**10.1**	5.5	3.5	2.7	9.2
N-REM (AB,ANES)	1.8	**53.5**	24.5	10.4	**4.9**	**3.1**	**2.5**	**9.0**
HMM	0.7	64.0	25.9	7.4	2.6	1.2	0.9	7.6
SC	3.7	52.4	23.5	11.0	5.9	2.7	4.5	9.9
FE	0.5	52.6	20.5	5.7	2.1	1.2	0.5	6.0

At the higher SNR levels, using the AB divergence does not have a considerable impact on the performance. The average WER between 0 dB and 20 dB slightly decreases from 5.2% to 5.0%. N-REM performs better than SC and HMM at 5 dB with a WER of 5.8% compared to 7.3% of HMM and 7.4% of SC. The best results at 5 dB and 10 dB are provided by the FE recognizer with WERs of 3.3% and 1.5%. At 15 dB and 20 dB, there is a performance gap between the N-REM and SC recognizers and the HMM and FE recognizers which benefit from the enhanced discriminative power of complex GMMs used in conjunction with MFCC features. N-REM performs better than SC with a WER of 3.5% at 15 dB and 2.3% at 20 dB compared to 4.8% and 4.5% of SC. The recognition performance on test set B is given in the lower panel of Table 7.1. In general, using the AB divergence does not have a noticeable influence in the case of mismatched noise.

7.4.2 CHIME-2

For the CHIME-2 data, the HMM recognizer uses speaker-dependent acoustic models trained on noisy data. These results are obtained using the HTK recognition toolkit and the details are available at the 2$^{\text{nd}}$ CHIME Challenge

Table 7.2: Keyword recognition accuracies in percentages obtained on the development and test set of the CHIME-2 database

(a) Development Set

SNR(dB)	-6	-3	0	3	6	9	*Avg*
N-REM (SP-KLD)	69.4	76.4	85.0	90.1	**92.9**	**93.3**	84.5
N-REM (SP-KLD,ANES)	70.4	77.9	84.8	90.4	92.6	93.8	85.0
N-REM (AB,ANES)	**75.4**	**78.8**	**86.3**	**90.5**	91.2	92.7	**85.8**
HMM	49.3	58.7	67.5	75.1	78.8	82.9	68.7
FE	68.0	72.2	80.9	86.7	89.0	90.5	81.2
HMM-FE	69.1	73.6	81.5	87.3	89.4	90.3	81.9
SC	75.5	81.4	87.5	89.9	92.4	92.3	86.5

(b) Test set

SNR(dB)	-6	-3	0	3	6	9	*Avg*
N-REM (SP-KLD)	69.3	76.8	84.5	**88.8**	**91.9**	**93.5**	84.1
N-REM (SP-KLD,ANES)	71.0	78.9	85.3	88.7	**91.9**	92.8	84.8
N-REM (AB,ANES)	**73.9**	**79.7**	**86.1**	88.0	90.9	92.6	**85.2**
HMM	49.7	57.9	67.8	73.7	80.8	82.7	68.8
FE	67.2	75.9	81.1	86.4	90.7	92.0	82.2
HMM-FE	67.0	77.0	81.8	87.0	91.2	92.4	82.7
SC	76.5	81.3	88.9	90.5	92.7	93.2	87.2

website[1]. The details of the SC, FE and HMM-FE recognition systems such as feature extraction schemes and dictionary sizes are described in [51]. The FE recognizer refers to the baseline NMF system trained on the reverberated data and HMM-regularized FE (HMM-FE) recognizer refers to the proposed system in [51]. The overcomplete dictionary used by SC, FE and HMM-FE recognizers contain 5000 speech and 5000 noise exemplars with exemplar length of 20 frames. The N-REM baseline without ANES uses 400 sniffed noise exemplars only which are extracted from the immediate context of the target utterances [184].

The keyword recognition accuracies (RA) obtained on the development and test sets of the CHIME-2 data are given in Table 7.2. The upper panel of each table presents the results provided by the baseline with and without ANES and the novel N-REM recognizer using the AB divergence with ANES. The lower panels list the results yielded by the comparable recognition systems. The highest performance gains are obtained at the lower SNR levels both for the

[1] `http://spandh.dcs.shef.ac.uk/chime_challenge/chime2013/chime2_task1.html`

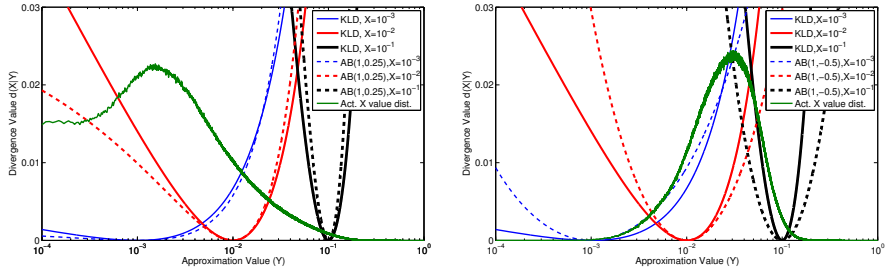
development and test set. The RAs obtained on the test set using the baseline without ANES are 69.3% at -6 dB, 76.8% at 0 dB and 84.5% at 3 dB. The second row of the upper panel presents the results provided by the baseline with ANES. The adaptive noise modeling technique improves the noise modeling capabilities providing recognition accuracies of 71.0%, 78.9% and 85.3% at the same SNR levels. Using the AB divergence with ANES, the recognition performance of the proposed setup further increases with RAs of 73.9% at -6 dB, 79.7% at 0 dB and 86.1% at 3 dB. The total absolute improvements at SNR levels of -6 dB, -3 dB and 0 dB are 4.6%, 2.9% and 1.6% respectively.

The RA of the proposed recognizer does not outperform the baseline setup at SNR levels of 6 dB and 9 dB. The mean RA increases from 84.5% to 85.8% on the development set and from 84.1% to 85.2% on the test set. The SC recognizer provides comparable performance with N-REM on the development set and slightly better performance on the test set with a RA of 76.5% at -6 dB, 88.9% at 0 dB and 93.2 at 9dB. The mean RA of the SC recognizer is 86.5% on the development set and 87.2 on the test set which is the best among all systems. The performance of FE and HMM-FE recognizers are similar to each other on both sets for all SNR levels and lower than the SC and N-REM recognizers.

7.5 Discussion

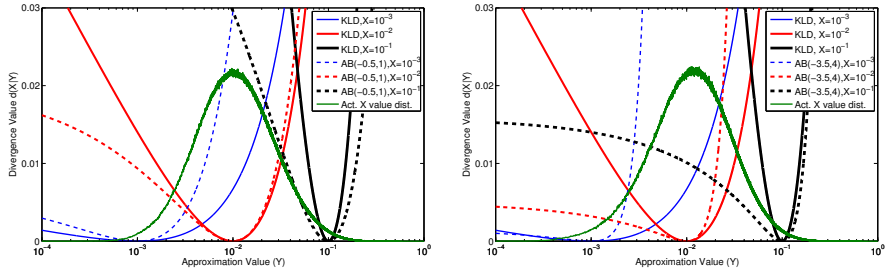
The results presented at the lower SNRs of test set A of AURORA-2 and both the test and development set of CHIME-2 demonstrate the improved noise robustness of N-REM using the AB divergence. The WER of 14.9% at the SNR level of -5 dB is the best published recognition performance to the best of our knowledge. The proposed recognizer picking an appropriate (α, β) value depending on the estimated SNR level performs an accurate speech and noise separation even at the lowest SNR levels. We discuss the reason for the performance gain by visualizing the behavior (regime) of the AB divergence for several (α, β) pairs and comparing them with the ones belonging to the generalized KLD.

Before elaborating on this issue, we revisit some system properties that have been mentioned in the earlier parts of the chapter which are relevant to the discussion. Firstly, N-REM uses l_2-normalized dictionaries where each individual time-frequency cell lies in the range of $[0, 1]$. The observation matrices are l_2-normalized similarly to the dictionaries and all values in the observation matrices are also in the range of $[0, 1]$. Secondly, the genuine room noise contaminating the CHIME-2 data has different statistical characteristics compared to the noise types in the test set A of AURORA-2 data. The former noise type has been



(a) AB divergence with $(1, 0.25)$ (best results for AURORA-2, clean)

(b) AB divergence with $(1, -0.5)$ (best results for AURORA-2, -5 dB)



(c) AB divergence with $(-0.5, 1)$ (best results for CHIME-2, 9 dB)

(d) AB divergence with $(-3.5, 4)$ (best results for CHIME-2, -6 dB)

Figure 7.5: Comparison of the divergence value $d(X|Y)$ between the AB divergence and generalized KLD for three observation time-frequency cell values $X = [0.001, 0.01, 0.1]$ and varying approximation values in the range of $0.0001 < Y < 1$. The green curves show the histogram of occurrence of the actual data values X on the respective data sets.

observed to be less stationary and more spectrotemporally diverse due to the various noise sources in the recording environment such as two adults, two children, TV, kitchen and laundry appliances, foot steps, toys, traffic, birds and so forth [19]. We can hence expect that the noise dictionaries for CHIME-2 provide a poorer match to the actual noise spectra compared to the case of AURORA-2. Moreover, the genuine room noise recordings contain reverberation as the recording environments have a $T_{60} = 0.3$ seconds.

Taking this information into account, we discuss the reasons of the performance gain by comparing the weighting and scaling behaviors of the generalized KLD and AB divergence with divergence parameters that provided the best results at the lowest and highest SNR level of both databases in Figure 7.5. The upper figures show the behavior of the AB divergence with $(1, 0.25)$ and $(1, -0.5)$

providing lower WERs on the clean test set and the test set at SNR level of -5 dB of AURORA-2 data respectively. We plot the divergence value for three different values of a time-frequency cell from an observation vector (denoted by X in the figure) with varying approximation values (denoted by Y in the figure). However, the recognition framework is invariant to scaling of the divergence, since scaling will not change the ranking of the recognition hypotheses. To avoid false interpretation of the divergence plots, we can therefore scale each AB divergence plot. We choose the scaling factor such that the local behavior of the AB divergence is the same as the local behavior of the KLD at a reference data value (X) of 0.01, i.e. KLD and AB divergence have the same curvature at the reference point. We choose a reference point of 0.01 because the probability density is high for all data sets considered. This can be verified from the distribution of real time-frequency cell values after the l_2-normalization as depicted in Figure 7.5. It is worth pointing out that the density plot belonging to the clean speech of AURORA-2 is obtained only from the segments that contain speech, i.e. silence frames are discarded.

Firstly, we observe that the AB divergence with $(1, 0.25)$ provides good accuracy for the clean test set of AURORA-2 data. From Figure 7.5a, it is clearly seen that the AB divergence downweights the smaller cells which contain little or no energy and puts more emphasis on the large cells, i.e. the spectral peaks. The very high SNR observed in the clean AURORA-2 data is reflected in the (green) density plot, where we observe a substantial fraction of the data with energy over 40 dB below the spectral peaks. To approximate such small spectral values accurately with a linear combination of atoms should not matter. Hence, there is no harm to reduce the penalty of underestimations of small values (red and blue curves below Y=0.01 in Figure 7.5a).

The noise robustness of N-REM depends on how well the noise exemplars model the actual noise conditions and how accurate the divergence weights the approximation error of time-frequency cells that define the characteristics of the noise source. This is vital for accurate separation of speech and noise. The best performance at SNR level of -5 dB of AURORA-2 data is obtained using the divergence parameters $(1, -0.5)$. The divergence with $(1, -0.5)$ is equally far from the generalized KLD (AB divergence with parameters $(1, 0)$) and the Itakura-Saito distance (AB divergence with parameters $(1, -1)$) on the $\alpha = 1$ line which is equal to the Beta divergence as a special case of the AB divergence. In this regard, it is a compromise between the generalized KLD and the Itakura-Saito distance which has been shown to be effective on source separation tasks using audio power spectrograms [42].

The behavior of the AB divergence with $(1, -0.5)$ is given in Figure 7.5b. From this figure, it can be seen that the divergence upweights the approximation errors of small time-frequency cells (e.g. X=0.001 in the figure), compared to

the generalized KLD. Looking at the data distribution (green), we see that this has no relevance, as actual data points in this order of magnitude are hardly observed. Medium (X=0.01) to large (X=0.1) data values should now be realized exactly during the approximation and both over and underestimation are penalized. Since the noise dictionaries are highly accurate in these tests, this is indeed a very good strategy to obtain an accurate signal decomposition in terms of speech and noise. Eventually, a few noise exemplars that resemble the actual noise component are selected from the noise dictionary providing an accurate separation and thus a high recognition performance.

This result matches up with the recognition results presented in [185] where a grid search on the AB plane has been performed to find the most appropriate divergence parameters at the lower SNRs of CHIME-2 data. The best recognition results are obtained on the $\alpha + \beta = 0.5$ line which passes through the $(1, -0.5)$ point. In case of CHIME-2 data, the recognition performance also benefits from using smaller α values on the $\alpha + \beta = 0.5$ line. The lower figures in Figure 7.5 show the behavior of the AB divergence with $(-0.5, 1)$ and $(-3.5, 4)$ providing best results on the test set at SNR level of 9 dB and -6 dB respectively. On CHIME-2 data, we observe that the optimal divergence choice downweights underestimations for small data values (dashed blue and red curves in the lower panes of Figure 7.5) compared to the AB divergence with $(1, -0.5)$ (dashed blue and red curves in Figure 7.5b). At the lower SNR (Figure 7.5d), this is even the case for the largest of spectral values (black curves). This is in line with missing data techniques for speech recognition [24], which constrain the clean speech model to be less than the observed noisy speech for time-frequency cells dominated by noise, while the clean speech model should be equal to the noisy observations for time-frequency cells dominated by speech. Since the noise dictionary is not expected to be very accurate for CHIME-2 data, while speech dictionaries are, the noise data is best explained with the speech exemplars, and a divergence metric reflecting the missing data approach is a sensible choice. At higher SNR (Figure 7.5c), the small time-frequency cells are most likely dominated by noise, so underestimation by the speech model should not be penalized. Large spectral values are most likely reliable and missing data theory then says that the speech prediction should match the noisy observation, which is indeed expressed by the black curve in Figure 7.5c. At low SNRs, large spectral values are often also dominated by noise and it is hence helpful to also allow not to penalize underestimation of large spectral values as well as expressed by the black dashed curve in Figure 7.5d.

After the discussion on the recognition performance in the matched noise scenario, we discuss the results obtained on the test set B of AURORA-2 data which contains mismatched noise types, i.e. noise types that are not available in the training data. Despite exemplar-based modeling being quite effective

in the case of matched noise, there is a performance gap between matched and mismatched noise scenarios for all exemplar-based techniques. The use of exemplars is most applicable in scenarios where the expected noise types can be predicted or when some noise exemplars can be readily obtained from the environment as in the CHIME-2 data. For N-REM, this gap is larger than for SC and FE since N-REM dictionaries contain smaller noise dictionaries. This results in a lower probability of having a suitable noise exemplar in the combined dictionaries with a similar spectral content with the unseen noise types.

Lastly, the performance of the adaptive noise modeling technique ANES has been investigated in a more realistic scenario in which noisy utterances are contaminated with genuine room noise. The results presented in Table 7.2 show that ANES combined with noise sniffing can model the genuine room noise more accurately compared to noise sniffing only. These results demonstrate the effective noise modeling of ANES achieved by picking a small number of noise exemplars which have similar characteristics to the actual noise contaminating the target utterance.

7.6 General Discussion and Concluding Remarks

In this chapter, we have aimed to improve the noise robustness of our noise robust exemplar matching approach by adopting the AB divergence which has a higher degree of freedom with two parameters compared to the generalized KLD. Various well-known distance/divergence measures such as the Euclidean distance, generalized Kullback-Leibler divergence, Itakura-Saito divergence and Hellinger distance are special cases of the AB divergence for different (α, β) values. By adjusting these parameters on the development data, the divergence is tailored for the representation of speech and noise for the best separation of the noisy mixtures. Applying the multiplicative update rules proposed for this new divergence in [20], the noisy utterances are modeled as a linear combination of exemplars that are organized in multiple dictionaries based on their duration and class. The presented recognition results have confirmed the improved noise robustness of the AB divergence compared to the conventional generalized KLD. After presenting the results, we have provided insight into the choice for the (α, β) pairs along the $\alpha + \beta = 0.5$ line depending on the parameters such as SNR and quality of the noise dictionaries.

Chapter 8

Speech Dictionary Design for N-REM using the AB-divergence

Exemplar-based acoustic modeling is based on labeled training segments that are compared with the unseen test utterances with respect to a dissimilarity measure. Using a larger number of accurately labeled exemplars provides better generalization thus improved recognition performance which comes with increased computation and memory requirements. We have recently developed a noise robust exemplar matching-based automatic speech recognition system which uses a large number of undercomplete dictionaries containing speech exemplars of the same length and label to recognize noisy speech. In this work, we investigate several speech exemplar selection techniques proposed for undercomplete speech dictionaries to find a trade-off between the recognition accuracy and the acoustic model size in terms of the amount of speech exemplars used for recognition. The exemplar selection criterion has be to chosen carefully as the amount of redundancy in these dictionaries is very limited compared to overcomplete dictionaries containing plenty of exemplars. The recognition accuracies obtained on the small vocabulary track of the 2^{nd} CHiME Challenge and the AURORA-2 database using the complete and pruned dictionaries are compared to investigate the performance of each selection criterion.

This chapter is adapted from: Emre Yılmaz, Jort F. Gemmeke and Hugo Van hamme, "*Data Selection for Noise Robust Exemplar Matching*", Submitted to INTERSPEECH 2015.

8.1 Introduction

Exemplar-based speech recognition systems [29, 36, 57, 70, 143, 159, 161] use labeled segments from training data to identify unseen speech. These approaches resemble the first attempts to solve the automatic speech recognition (ASR) problem performing dynamic time warping [134, 144, 178]. The recognition can be performed by comparing these labeled segments with the segments from the test utterances with respect to a dissimilarity measure. Though exemplars provide the most natural duration and trajectory modeling when compared to its statistical counterparts, e.g. hidden Markov models (HMM) or deep neural networks (DNN), large amounts of data are required to handle the acoustic variation among different utterances.

In order to reduce high memory and computational power requirements, several exemplar selection algorithms are proposed in [149, 158]. The main goal of these techniques is to remove less informative exemplars, e.g. duplicates or rarely used ones, or whose presence result in inaccurate recognition and achieve comparable recognition accuracies using only a portion of the exemplars. Statistical acoustic model training also benefits from data selection as the training times are reduced significantly and sometimes the recognition performance is improved due to the reduced noise and redundancy in the training data [63, 91, 125, 179].

Using exemplars in a sparse representation (SR) formulation provides significantly improved noise robustness and exemplar-based sparse representations have been successfully used for feature extraction, speech enhancement and noise robust speech recognition tasks [50, 84, 142, 162]. These approaches model the acoustics using fixed length exemplars which are labeled at frame level and stored in the columns of a single overcomplete dictionary. Noisy speech segments are jointly approximated as a sparse linear combination of speech and noise exemplars with exemplar weights obtained by solving a regularized convex optimization problem.

Reducing the dimensions of large datasets stored in a single overcomplete dictionary has been investigated in different fields and several matrix decompositions such as the singular value decomposition (SVD), rank revealing QR decomposition, CUR matrix decomposition, interpolative decomposition (ID) have been used to obtain a low-rank matrix approximation of the complete data matrix [64]. Although the SVD is known to provide the best rank-k approximation, interpretation of the principal components is difficult in data analysis [115]. Therefore, several CUR matrix decompositions have been proposed in which a matrix is decomposed as a product of three matrices $\mathbf{C}$, $\mathbf{U}$, $\mathbf{R}$ and the matrices $\mathbf{C}$ and $\mathbf{R}$ consist of a subset of the actual columns and rows respectively [38, 44, 61]. Several computationally efficient exemplar selection

techniques are introduced and applied to polyphonic music transcription task using an overcomplete dictionary containing exemplars of different musical notes in [5]. [85] discusses various ways of reducing the speech and noise dictionaries for an exemplar-based sparse representations approach applied to a noise robust ASR task.

In this chapter, we focus on the noise robust exemplar matching (N-REM) framework [184] which is an exemplar matching recognition system with noise modeling capabilities. In this framework, the recognizer uses different length exemplars organized in separate dictionaries based on their duration and label (the associated speech unit) [184]. The input speech segments are approximated in a sparse representations formulation, i.e. as a linear combination of the exemplars in each dictionary. Compared to a system using fixed-length exemplars stored in a single dictionary, using separate dictionaries for each class provides better classification as input speech segments are approximated as a combination of exemplars belonging to the same class only. Moreover, each exemplar is associated with a single speech unit and the natural duration distribution of each speech unit in the training data is preserved yielding exemplars of different lengths. This recognizer adopts a reconstruction error based back-end, i.e. the recognition is performed by comparing the approximation quality for different classes quantified by a divergence measure and choosing the class sequence that minimizes the total reconstruction error. In [186], we have proposed to use the alpha-beta divergence [20] in place of the generalized Kullback-Leibler divergence which has been shown to be more robust against background noise.

The exemplar selection techniques discussed in this chapter differ from previous work as the dictionaries store a lot less exemplars due to the use of multiple dictionaries for each exemplar length and label. Compared to the overcomplete dictionaries with a large number of data points, the redundancy in the undercomplete dictionaries used by N-REM is quite limited. Therefore, removing a few informative data points may already result in significant decreases in the recognition accuracy. We have presented the initial findings of our efforts to select a subset of speech exemplars in [182] and reported some promising recognition results on a clean digit recognition task. In this work, we extend the investigation of the proposed exemplar selection technique with the best performance, namely *collinearity reduction*, on all available SNR levels of the small vocabulary track of the 2$^{\text{nd}}$ CHiME Challenge and the AURORA-2 database. Moreover, in addition to this technique, we propose a symmetric AB-divergence-based k-medoids algorithm for exemplar selection from undercomplete dictionaries. The AB-divergence is chosen as a dissimilarity measure to be consistent with the recognition setup.

8.2 Noise Robust Exemplar Matching

Training frame sequences representing various speech units (speech exemplars) are extracted based on the state-level alignments obtained using a conventional HMM-based recognizer. Speech exemplars, each comprised of D mel frequency bands and spanning l frames, are reshaped into a single vector and stored in the columns of a speech dictionary $\mathbf{S}_{c,l}$: one for each class c and each frame length l. Each dictionary is of dimensionality $Dl \times N_{c,l}$ where $N_{c,l}$ is the number of available speech exemplars of class c and frame length l. Similarly, a noise dictionary $\mathbf{N}_l$ for each frame length l is formed by reshaping the noise exemplars. Each speech dictionary is concatenated with the noise dictionary of the same length to form a combined dictionary $\mathbf{A}_{c,l} = [\mathbf{S}_{c,l}\, \mathbf{N}_l]$ of dimensionality $Dl \times M_{c,l}$ where $M_{c,l}$ is the total number of available speech and noise exemplars.

An observed noisy (and/or reverberated) speech segment of frame length T frames is also reshaped into vectors by applying a sliding window approach [50] with window length of l frames and stored in an observation matrix $\mathbf{Y}_l = [\mathbf{y}_l^1, \mathbf{y}_l^2 ..., \mathbf{y}_l^{(T-l+1)}]$ of dimensionality $Dl \times (T - l + 1)$. Due to multiple-length exemplars, the window length l is varied between the minimum exemplar length $l_{\min}$ and maximum exemplar length $l_{\max}$ yielding observation matrices $\mathbf{Y}_l$ for $l_{\min} \leq l \leq l_{\max}$. For every class c, each observation vector $\mathbf{y}_l$ is expressed as a linear combination of the exemplars that are stored in the dictionaries of the same length: $\mathbf{y}_l \approx \sum_{m=1}^{M_{c,l}} x_{c,l}^m \mathbf{a}_{c,l}^m = \mathbf{A}_{c,l}\mathbf{x}_{c,l}$ for $x_{c,l}^m \geq 0$. Here, $\mathbf{x}_{c,l}$ is an $M_{c,l}$-dimensional non-negative weight vector. The combined dictionaries consisting of speech and noise exemplars are presumed to model all acoustic variability in the observed signal due to pronunciation variation, background noise and so forth. This model can also model reverberation by storing reverberated speech exemplars rather than clean speech exemplars.

The exemplar weights $\mathbf{x}_{c,l}$ are obtained by minimizing the cost function consisting of a single term which quantifies the approximation error $d(\mathbf{y}_l, \mathbf{A}_{c,l}\mathbf{x}_{c,l})$ for non-negative exemplar weights. This optimization problem can be solved with non-negative sparse coding (NSC) [79]. The value of approximation error is highly dependent on the divergence measure d and the representation of speech and noise sources. Motivated by its capabilities to weight and scale the individual ratios of the noisy speech and its approximation, $\mathbf{y}_l^i / \hat{\mathbf{y}}_{c,l}^i$ where $\hat{\mathbf{y}}_{c,l} = \mathbf{A}_{c,l}\mathbf{x}_{c,l}$, the AB divergence is used for d. The AB divergence $d_{AB}^{(\alpha,\beta)}(\mathbf{y}, \hat{\mathbf{y}})$

is defined as

$$= \begin{cases} -\frac{1}{\alpha\beta}\sum_{k=1}^{K}\left(y_k^{\alpha}\hat{y}_k^{\beta} - \frac{\alpha}{\gamma}y_k^{\gamma} - \frac{\beta}{\gamma}\hat{y}_k^{\gamma}\right) & \text{for} \quad \alpha,\beta,\gamma \neq 0, \\ \frac{1}{\alpha^2}\sum_{k=1}^{K}\left(y_k^{\alpha}\log(\frac{y_k^{\alpha}}{\hat{y}_k^{\alpha}}) - y_k^{\alpha} + \hat{y}_k^{\alpha}\right) & \text{for} \quad \alpha \neq 0, \beta = 0 \\ \frac{1}{\alpha^2}\sum_{k=1}^{K}\left(\log(\frac{\hat{y}_k^{\alpha}}{y_k^{\alpha}}) + \frac{y_k^{\alpha}}{\hat{y}_k^{\alpha}} - 1\right) & \text{for} \quad \alpha = -\beta \neq 0, \\ \frac{1}{\beta^2}\sum_{k=1}^{K}\left(\hat{y}_k^{\beta}\log(\frac{\hat{y}_k^{\beta}}{y_k^{\beta}}) - \hat{y}_k^{\beta} + y_k^{\beta}\right) & \text{for} \quad \alpha = 0, \beta \neq 0, \\ \frac{1}{2}\sum_{k=1}^{K}\left(\log(y_k) - \log(\hat{y}_k)\right)^2 & \text{for} \quad \alpha,\beta = 0 \end{cases} \tag{8.1}$$

where $\gamma = \alpha+\beta$. The two parameters of the AB divergence can be automatically adjusted based on the amount of contamination in the target utterance as the recognition performance for different noise levels depends on the emphasized (reliable) time-frequency bins. For the NSC solution, we apply the multiplicative update rule minimizing the approximation error $d(\mathbf{y}_l, \mathbf{A}_{c,l}\mathbf{x}_{c,l})$ using the AB divergence for $\alpha \neq 0$ which is given in [186].

All observation matrices Y_l for $l_{\min} \leq l \leq l_{\max}$ are approximated using the combined dictionaries $\mathbf{A}_{c,l}$ of the corresponding length by applying the multiplicative update rule. To quantify the approximation quality, we use the reconstruction error between the noisy speech segments and their approximations. The multiplicative update rule is applied iteratively until the reconstruction error provides enough discrimination between different classes. The number of iterations that satisfies this criterion has been investigated in pilot experiments. After a fixed number of iterations for all dictionaries, the reconstruction errors between the observation matrix $\mathbf{Y}_l$ and its approximations $\mathbf{A}_{c,l}\mathbf{x}_{c,l}$ are calculated for $l_{\min} \leq l \leq l_{\max}$. As the label of each dictionary is known, decoding is performed by applying dynamic programming [128] to find the class sequence that minimizes the reconstruction error (taking the grammar into account if necessary).

8.3 Exemplar Selection Techniques

The N-REM recognition scheme benefits from discarding redundant speech exemplars due to two main reasons. First, the computational load mainly due to the iterative evaluation of the multiplicative update rule reduces proportional to the dictionary sizes. Furthermore, the memory required to store the pruned dictionaries is much less than storing the complete dictionaries. For this purpose, we investigate the impact of two exemplar selection methods, namely collinearity

reduction and k-medoids with symmetric AB divergence, on the recognition accuracy in both clean and noisy conditions.

8.3.1 Collinearity Reduction (CR)

The CR selection technique discards exemplars that are well approximated by the other exemplars of the same length and class (i.e. other exemplars in the same dictionary). The exemplars with larger reconstruction errors are expected to contribute more when approximating unseen noisy segments compared to the ones with smaller reconstruction errors. Therefore, the CR technique compares the reconstruction errors for all exemplars in a dictionary by approximating each exemplar as a linear combination of the other exemplars in the same dictionary with non-negative weights. This idea is applied iteratively by removing the exemplar that is approximated with the minimum reconstruction error at each iteration until the minimum number of exemplars requirement in a dictionary is met.

8.3.2 K-medoids with AB Divergence (KMED)

The KMED selection technique is based on the well-known k-medoids technique, PAM [94], using a symmetric version of the AB divergence as a novel dissimilarity measure. The symmetric version of the AB divergence given in Equation (8.1) is obtained as $\frac{1}{2}\left[d_{AB}^{(\alpha,\beta)}(\mathbf{y},\hat{\mathbf{y}}) + d_{AB}^{(\alpha,\beta)}(\hat{\mathbf{y}},\mathbf{y})\right]$. The higher computational complexity of the PAM technique mentioned in [65] is not valid in this scenario as the number of speech exemplars in each dictionary is mostly on the order of magnitude one and two. This selection technique is applied to every dictionary to obtain a certain number of medoids that are expected to represent the convex hull formed by the complete dictionary accurately enough. The divergence parameters are chosen based on the recognition performance of the speech dictionaries on clean speech and the ones providing the best clean speech recognition performance are used during the exemplar selection.

8.4 Experimental Setup

8.4.1 Databases

The training material of AURORA-2 [77] consists of a clean and a multi-condition training set, each containing 8440 utterances. The multi-condition training set

was constructed by mixing the clean utterances with noise at SNR levels of 20, 15, 10 and 5 dB. Test set A and B consist of 4 clean and 24 noisy datasets at six SNR levels between -5 and 20 dB. The noise types of test set A match the multi-condition training set. Each subset contains 1001 utterances with one to seven digits 0-9 or oh. To reduce the simulation times, we subsampled the test sets by a factor of 4 (1000 utterances per SNR).

The small vocabulary track of the 2$^{\text{nd}}$ CHiME Challenge [172] addresses the problem of recognizing commands in a noisy and reverberant living room. The clean utterances contain utterances from 34 speakers reading 6-word sequences of the form *command-color-preposition-letter-digit-adverb*. There are 25 different letters, 10 different digits and 4 different alternatives for each of the other classes. The recognition accuracy of a system is calculated based on the correctly recognized letter and digit keywords.

8.4.2 Dictionary Creation and Implementation Details

The speech exemplars of AURORA-2 data are extracted from the clean training set. Acoustic feature vectors are represented in mel-scaled magnitude spectra with 23 frequency bands. The speech exemplars representing half-digits are segmented by a conventional HMM-based system. The complete dictionary contains in total 52,295 speech exemplars excluding 990 silence exemplars. The number of noise exemplars varies depending on the duration of the noise-only sequences that are selected by ANES. On average, the recognizer with the pruned dictionaries containing 20% of the exemplars in each dictionary uses 11,355 and 1,044 noise exemplars/utterance in total at SNR level of -5 dB and clean speech respectively. The divergence parameters (α,β) for the KMED selection technique are set to 1 and 0.25 respectively. The minimum and maximum exemplar lengths are 8 and 40 frames respectively. Exemplars longer than 40 frames are omitted to limit the number of dictionaries. The recognizer uses 675 dictionaries in total. In the recognition phase, noise dictionaries are created by performing noise sniffing and active noise exemplar selection [184]. The combined dictionaries and observation matrices are l_2-normalized for all SNR levels. The multiplicative update rule is iterated 100 times for convergence of all frame lengths. The further details are given in [186]. The word error rate (WER) has been used to quantify the recognition accuracy for the AURORA-2 digit recognition task.

The exemplars and noisy speech segments from CHIME-2 data are represented as mel-scaled magnitude spectral features extracted with a 26 channel mel-scaled filter bank ($D = 26$). The frame length is 25 ms and the frame shift is 10 ms. The binaural data is averaged in the spectral domain to obtain 26-dimensional

feature vectors. The exemplars are extracted from the reverberated utterances in the training set according to the state-based segmentations obtained using the acoustic models in the toolkit provided with the database. Exemplars belonging to each speaker are organized in separate dictionary sets for speaker-dependent modeling yielding 34 different dictionary sets. Based on the availability of the exemplars, the minimum and maximum exemplar lengths are 4 and 40 frames respectively. Half-word exemplars seemed to generalize sufficiently to unseen data for the recognition task. Dictionary sizes vary with different classes and speakers. The divergence parameters (α,β) for the KMED selection technique are set to 1 and 0 respectively. *Prewarping* [183] is applied to boost the modeling capabilities of the underpopulated speech dictionaries (especially for the ones belonging to letters due to the high number of alternatives and hence the small number of exemplars per class) and it is limited to a single frame. The number of exemplars in each dictionary after prewarping is limited to 50. The noise modeling is detailed in [186]. The multiplicative update rule is iterated 25 times to obtain the exemplar weights. The columns of the combined dictionaries and observation matrices are l_2-normalized. The further details are given in [186]. The keyword recognition accuracy (RA) is used to evaluate the system performance on the CHIME-2 data.

8.5 Results and Discussion

The exemplar selection techniques described in Section 8.3 are applied to the speech dictionaries obtained from AURORA-2 and CHIME-2 data and the recognition performance of the recognizers using only 20% of the exemplars per dictionary are presented in Table 8.1 and 8.2. The results obtained using conventional multi-condition trained GMM/HMM and other exemplar-based sparse representation systems, namely sparse classification (SC) and feature enhancement (FE), are also provided for comparison. The details of these systems are available in [50]. The baseline results obtained with the complete dictionaries and the best results provided by the pruned dictionaries are given in bold.

A pruning rate of 80%, i.e. using 20% of the exemplars in a dictionary, is chosen based on the initial results presented in [182]. This choice aims to compare the amount of degradation in the recognition accuracy when pruning goes further than the *safe* pruning rate of 70% which is defined as the largest pruning rate without significant recognition accuracy loss [182]. We compare the CR and KMED techniques with the CUR decomposition which is a randomized column selection algorithm proposed as a part of the CUR matrix decomposition in [115]. This algorithm randomly selects a subset of the columns of a data

Table 8.1: Word error rates in % obtained on test set A and B of AURORA-2 using 20% of exemplars in each dictionary

(a) Test set A

SNR(dB)	clean	-5	0	5	10	15	20	*0-20*
N-REM	**1.8**	**14.9**	**8.5**	**5.8**	**4.7**	**3.5**	**2.3**	**5.0**
CR	**2.8**	19.8	**10.8**	8.0	**6.3**	**4.7**	**3.5**	**6.7**
KMED	3.0	**18.6**	10.9	**7.9**	6.4	5.0	4.1	6.9
CUR	4.1	20.2	12.6	9.0	7.4	5.6	4.5	7.8
RND	4.1	20.4	12.5	9.0	7.2	5.5	4.5	7.8
GMM	0.7	60.8	24.3	7.3	2.9	1.3	0.8	7.3
SC	3.7	35.2	13.8	7.4	5.6	4.8	4.5	7.2
FE	0.5	30.4	10.7	3.3	1.5	1.1	0.7	3.5

(b) Test set B

SNR(dB)	clean	-5	0	5	10	15	20	*0-20*
N-REM	**1.8**	**53.5**	**24.5**	**10.4**	**4.9**	**3.1**	**2.5**	**9.0**
CR	**2.8**	56.7	27.5	12.5	7.0	**4.7**	**3.5**	11.0
KMED	3.0	58.5	**25.9**	**11.7**	**6.9**	5.0	4.6	**10.8**
CUR	4.1	57.6	26.4	13.0	7.4	5.7	5.1	11.5
RND	4.1	**56.1**	26.9	12.8	7.2	5.6	4.3	11.4
GMM	0.7	64.0	25.9	7.4	2.6	1.2	0.9	7.6
SC	3.7	52.4	23.5	11.0	5.9	2.7	4.5	9.9
FE	0.5	52.6	20.5	5.7	2.1	1.2	0.5	6.0

matrix with respect to the probability distribution computed as the normalized statistical leverage scores. The CUR decomposition has been successfully applied in selecting a very small number of exemplars from an overcomplete dictionary without a significant recognition accuracy loss. We further provide the recognition accuracies obtained using the randomly pruned dictionaries (RND).

The WERs obtained on the clean test set of AURORA-2 are presented in the middle panel of Table 8.1a and 8.1b. The N-REM performance using the complete dictionaries is given in the first row of the tables. The clean speech performance of CR is the best among the results obtained with the pruned dictionaries with a WER of 2.8% compared to 1.8% yielded by the complete dictionaries. KMED also provides a comparable result with a WER of 3.0%. These results are consistent with the clean speech recognition results of CR

Table 8.2: Keyword recognition accuracies in % obtained on the dev. and test set of CHIME-2 using 20% of exemplars in each dictionary

(a) Development Set

SNR(dB)	-6	-3	0	3	6	9	*Avg*
N-REM	**75.4**	**78.8**	**86.3**	**90.5**	**91.2**	**92.7**	**85.8**
CR	71.5	77.7	83.6	90.0	90.6	92.3	84.3
KMED	**73.0**	**77.8**	**84.7**	**90.3**	**91.3**	**92.4**	**84.9**
CUR	69.3	76.3	82.3	87.9	89.7	91.9	82.9
RND	70.4	76.1	81.8	88.8	89.2	91.5	83.0
GMM	49.3	58.7	67.5	75.1	78.8	82.9	68.7
FE	68.0	72.2	80.9	86.7	89.0	90.5	81.2
HMM-FE	69.1	73.6	81.5	87.3	89.4	90.3	81.9
SC	75.5	81.4	87.5	89.9	92.4	92.3	86.5

(b) Test set

SNR(dB)	-6	-3	0	3	6	9	*Avg*
N-REM	**73.9**	**79.7**	**86.1**	**88.0**	**90.9**	**92.6**	**85.2**
CR	**72.1**	**78.7**	**84.9**	**87.1**	**90.6**	**91.8**	**84.2**
KMED	71.8	77.9	83.8	86.9	89.4	91.6	83.6
CUR	70.1	77.4	82.9	85.5	88.7	90.4	82.5
RND	70.6	77.3	82.9	86.0	88.6	90.5	82.7
HMM	49.7	57.9	67.8	73.7	80.8	82.7	68.8
FE	67.2	75.9	81.1	86.4	90.7	92.0	82.2
HMM-FE	67.0	77.0	81.8	87.0	91.2	92.4	82.7
SC	76.5	81.3	88.9	90.5	92.7	93.2	87.2

presented in [182]. Dictionaries pruned with the other techniques yield worse performance.

The results on the noisy sets of test set A are given in the rightmost panel of Table 8.1a. These results further demonstrate the effectiveness of CR and KMED in the noisy scenarios. N-REM with complete dictionaries has a WER of 5.0% on average. CR and KMED provide a WER of 6.7% and 6.9% respectively. CUR performs as poor as RND on this exemplar selection task yielding a WER of 7.8%. The results on test set B, which are presented in Figure 8.1b, show a similar trend and the best results in the mismatched noise case are obtained using the dictionaries pruned by CR at high SNR levels and by KMED at low SNR levels. At -5 dB of test set B, RND provides the best results which is

explained by the minor impact of the speech dictionaries on the recognition accuracy due to very poor noise modeling. CR and KMED perform better than CUR and RND on average similar to the matched noise case.

The RAs obtained on the development and test sets of CHIME-2 data are shown in Table 8.2. On the development set, KMED and CR yield an average RA of 84.9% and 84.3% compared to 85.8% of the N-REM baseline. CUR and RND have a comparable RA of 82.9% and 83.0% respectively. On the test set, CR provides an average RA of 84.2% which is slightly better than 83.6% of KMED. These results are higher than 82.5% of CUR and 82.7% of RND.

From these results, it can be concluded that the CR and KMED techniques achieve effective exemplar selection from undercomplete dictionaries by reducing the dictionary sizes significantly without a significant loss in the recognition performance, especially at higher SNR levels. Based on the geometrical interpretation of this exemplar selection task as explained in [182], these techniques pick the exemplars that preserve the convex hulls formed by the speech dictionaries in the positive orthant. As a result, the dictionaries pruned by CR and KMED have a more precise description of each speech unit in the high-dimensional feature space compared to the other techniques and the noisy mixtures can still be separated accurately by picking a low number of noise and speech exemplars with much less computational and memory requirements compared to the complete dictionaries.

8.6 Conclusion

This chapter investigates the performance of several exemplar selection approaches proposed for picking the most informative exemplars from undercomplete dictionaries which are used in the noise robust exemplar matching framework. We first apply the *collinearity reduction* approach, which has shown superior performance on clean speech in previous work, to noisy speech to explore how robust the pruned dictionaries against background noise. Furthermore, we investigate the performance of a k-medoids exemplar selection approach which uses a novel dissimilarity measure, namely the symmetric alpha-beta divergence, in accordance with the recognizer. The dictionaries pruned by both techniques have performed considerably better than random pruning and the column selection of the CUR decomposition which has provided impressive results on overcomplete dictionaries.

Chapter 9

Speech Enhancement Using N-REM

In this chapter, we propose a single-channel speech enhancement system based on the noise robust exemplar matching (N-REM) framework using coupled dictionaries. N-REM approximates noisy speech segments as a sparse linear combination of speech and noise exemplars that are stored in multiple dictionaries based on their length and associated speech unit. The dictionaries providing the best approximation of the noisy mixtures are used to estimate the speech component. We further employ a coupled dictionary approach that performs the approximation in the lower dimensional mel domain to benefit from the reduced computational load and better generalization, and the enhancement in the short-time Fourier transform (STFT) domain for higher spectral resolution. The proposed enhancement system is shown to have superior performance compared to the exemplar-based sparse representations approach using fixed-length exemplars in a single overcomplete dictionary.

This chapter is adapted from: Emre Yılmaz, Deepak Baby and Hugo Van hamme, "*Noise Robust Exemplar Matching with Coupled Dictionaries for Single-Channel Speech Enhancement*", Submitted to EUSIPCO 2015.

9.1 Introduction

Single-channel speech enhancement approaches aim to reduce the amount of background noise in speech signals recorded by a microphone and improve the speech intelligibility and quality. These techniques can also be used in the front end of other speech processing tasks such as automatic speech recognition (ASR) to alleviate the degradation due to the background noise. Denoising of monaural speech data is still a rather challenging problem even after the intensive research over several decades [112]. Numerous statistical and data-driven approaches have been proposed to tackle the problem [62, 89, 116, 117, 119, 151, 154, 174] (and references therein).

This chapter presents a novel exemplar-based speech enhancement approach, dubbed *noise robust exemplar matching* (N-REM), which performs denoising using the actual occurrences of speech and noise extracted from training data. Unlike previous exemplar-based sparse representations (SR) of speech using fixed-length exemplars in a single overcomplete dictionary [12, 50, 84, 92, 118, 153, 162], the proposed approach uses exemplars of multiple lengths, each associated with a single speech unit such as phones, syllables, half-words or words [29, 36, 57]. These exemplars are organized in multiple dictionaries based on the their length and class (associated speech unit). Using separate dictionaries for different speech units is motivated by the geometrical interpretation of SR-based source separation. It is known that the larger the distance between the convex hull of the basis vectors belonging to speech and noise sources are, the better the separation is [37]. Hence, the use of separate dictionaries for each speech unit provides a more precise representation in the high-dimensional feature space.

Previously, the N-REM framework has been shown to perform reasonably well on small vocabulary ASR tasks [184]. This chapter describes the initial efforts towards an N-REM based speech enhancement approach. In addition, we incorporate a coupled dictionaries approach [12] which uses a front-end dictionary containing lower dimensional features to obtain the decomposition, and a back-end dictionary containing the full-resolution spectral representations to reconstruct the speech and noise sources. In this way, the proposed approach benefits from the advantages of the lower dimensional features like better generalization and lower computational complexity during the decomposition and higher spectral resolution during the reconstruction of the speech component. For a reliable reconstruction, the mapping between the corresponding exemplars in both the dictionaries should be one-to-one which is realized by extracting the corresponding exemplars of the coupled dictionaries jointly from the same piece of training data.

9.2 Noise Robust Exemplar Matching

9.2.1 Exemplar extraction and dictionary creation

Training frame sequences associated with a single speech unit (speech exemplars) are extracted based on the state-level alignments obtained using a conventional HMM-based recognizer. Every speech exemplar is represented both in the full-resolution spectral domain (henceforth STFT exemplars) with K frequency bins and lower dimensional mel domain (henceforth mel exemplars) with D mel frequency bands. For the transformation between these domains, we use a STFT-to-mel matrix, $\mathbf{C}$, of dimensionality $D \times K$ which contains the vectorized magnitude response of D mel bands in its rows.

Mel speech exemplars, each comprised of D mel frequency bands and spanning l frames, are reshaped into a single vector and stored in the columns of a mel speech dictionary $\mathbf{S}^M_{c,l}$: one for each class c and each length l. Each dictionary is of dimensionality $Dl \times R_{c,l}$ where $R_{c,l}$ is the number of available mel speech exemplars of class c and length l. Similarly, a mel noise dictionary $\mathbf{N}^M_l$ for each length l is formed by reshaping the noise exemplars. Each mel speech dictionary is concatenated with the mel noise dictionary of the same length to form a combined mel dictionary $\mathbf{A}^M_{c,l} = [\mathbf{S}^M_{c,l}\, \mathbf{N}^M_l]$ of dimensionality $Dl \times P_{c,l}$ where $P_{c,l}$ is the total number of available speech and noise exemplars. The same procedure is followed using the STFT speech and noise exemplars to obtain the combined STFT dictionaries $\mathbf{A}^F_{c,l} = [\mathbf{S}^F_{c,l}\, \mathbf{N}^F_l]$ of dimensionality $Kl \times P_{c,l}$.

9.2.2 Decomposition of noisy speech

The decomposition of noisy mixtures into speech and noise components is performed only in the mel domain. Every observed noisy speech segment of length T frames is also reshaped into vectors by applying a sliding window approach [50] with window length of l frames and stored in an observation matrix $\mathbf{Y}_l = [\mathbf{y}^1_l, \mathbf{y}^2_l ..., \mathbf{y}^{(T-l+1)}_l]$ of dimensionality $Dl \times (T - l + 1)$. Due to multiple-length exemplars, the window length l is varied between the minimum exemplar length $l_{\min}$ and maximum exemplar length $l_{\max}$ yielding observation matrices $\mathbf{Y}_l$ for $l_{\min} \leq l \leq l_{\max}$. For every class c, each observation vector $\mathbf{y}_l$ is expressed as a linear combination of the exemplars that are stored in the dictionaries of the same length:

$$\mathbf{y}_l \approx \sum_{p=1}^{P_{c,l}} x^p_{c,l} \mathbf{a}^{M,p}_{c,l} = \mathbf{A}^M_{c,l} \mathbf{x}_{c,l} \qquad \text{s.t.} \qquad x^p_{c,l} \geq 0 \tag{9.1}$$

where $\mathbf{x}_{c,l}$ is an $P_{c,l}$-dimensional non-negative weight vector. The sparse solutions of $\mathbf{x}_{c,l}$ yield a more realistic approximation of the observed segments without overfitting and have been shown to provide better recognition results [80, 174].

The combined dictionaries consisting of speech and noise exemplars are presumed to model all acoustic variability in the observed signal due to pronunciation variation, background noise and so forth. This model can also cope with reverberation by storing reverberated speech exemplars rather than clean speech exemplars.

9.2.3 Obtaining the exemplar weights

The non-negative exemplar weights $\mathbf{x}_{c,l}$ are obtained by minimizing the cost function,

$$d(\mathbf{y}_l, \mathbf{A}^M_{c,l}\mathbf{x}_{c,l}) + \sum_{p=1}^{P_{c,l}} x^p_{c,l}\Lambda_p \qquad \text{s.t.} \qquad x^p_{c,l} \geq 0 \tag{9.2}$$

where $\mathbf{\Lambda}$ is an $P_{c,l}$-dimensional vector. The first term is the divergence between the observation vector and its approximation. The second term is a regularization term which penalizes the l_1-norm of the weight vector to produce a sparse solution. $\mathbf{\Lambda}$ contains non-negative values and controls how sparse the resulting vector $\mathbf{x}$ is. Defining $\mathbf{\Lambda}$ as a vector, the amount of sparsity enforced on different types of exemplars can be adjusted. In this work, the regularized optimization problem with the cost function in Equation (9.2) is solved by applying non-negative sparse coding (NSC) [79]. The generalized KLD is used for d which is commonly used in source separation problems and shown to produce better results than Euclidean distance when used in conjunction with mel-scaled spectral features [174],

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k. \tag{9.3}$$

All observation matrices $\mathbf{Y}_l$ for $l_{\min} \leq l \leq l_{\max}$ are approximated using the combined mel dictionaries $\mathbf{A}^M_{c,l}$ of the corresponding length by applying the multiplicative update rule given in [184]. To quantify the approximation quality, we use the reconstruction error between the noisy speech segments and their approximations. After a fixed number of iterations for all dictionaries, the reconstruction errors between the observation matrix $\mathbf{Y}_l$ and its approximations $\mathbf{A}^M_{c,l}\mathbf{x}_{c,l}$ are calculated for $l_{\min} \leq l \leq l_{\max}$. As the label of each dictionary is known, decoding is performed by applying dynamic programming to find

the class sequence that minimizes the reconstruction error to find the best approximation of the target utterance.

9.2.4 Speech enhancement

After finding the best matching dictionaries, the denoising is performed in two ways, either reconstructing the speech and noise components in mel or STFT domain. The former approach provides the frame-wise mel speech and noise estimates, $\hat{s}^M_{c,l}$ and $\hat{n}^M_{c,l}$, that are obtained after removing the windowing effect by adding the components belonging to overlapping windows from the estimates $S^M_{c,l}X^s_{c,l}$ and $N^M_l X^n_{c,l}$ respectively. Here, $X^s_{c,l}$ refers to the exemplar weights of the speech exemplars and $X^n_{c,l}$ refers to the exemplar weights of the noise exemplars. The frame-level Wiener-like filter is then obtained as in [12],

$$W = \mathbf{C}^T \hat{s}^M_{c,l} \oslash (\mathbf{C}^T(\hat{s}^M_{c,l} + \hat{n}^M_{c,l})) \tag{9.4}$$

Since $\mathbf{C}$ contains triangular shaped filter-banks, this extrapolation is the same as the piece-wise linear interpolation between D points (mel bands) spread across the 1 to K frequency bins. The resulting filters always fall in the D-dimensional subspace defined by the columns of $\mathbf{C}^T$ which cannot account for all the added noise content along the K dimensional DFT space. The enhanced speech obtained after applying this filter on the noisy DFT thus will result in a sub-optimal noise suppression.

The coupled dictionary approach remedies this problem by using the STFT speech and noise dictionaries to obtain the the frame-wise speech and noise estimates $\hat{s}^F_{c,l}$ and $\hat{n}^F_{c,l}$ from the estimates $S^F_{c,l}X^s_{c,l}$ and $N^F_l X^n_{c,l}$ respectively. The resulting Wiener-like filter can be written as

$$W_{cd} = \hat{s}^F_{c,l} \oslash (\hat{s}^F_{c,l} + \hat{n}^F_{c,l}). \tag{9.5}$$

The complex spectrogram of the enhanced signal is obtained by combining the enhanced magnitude spectrogram with the phase information obtained from the noisy speech. The speech signal in the time domain is obtained using the overlap-add method.

9.3 Experimental Setup

The enhancement performance of N-REM is evaluated on the test set A and B of the AURORA-2 corpus [77]. The training material of AURORA-2 consists of a clean and a multi-condition training set, each containing 8440 utterances

with one to seven digits in American English. The multi-condition training set was constructed by mixing the clean utterances with noise at SNR levels of 20, 15, 10 and 5 dB. Test set A consists of 4 clean and 24 noisy datasets with four noise types (subway, babble, car and exhibition) at six SNR levels, 20, 15, 10, 5, 0 and -5 dB. The noise types of this test set match the multi-condition training set. Test set B has the same number of test sets with four different noise types (restaurant, street, airport, station) at the same SNR levels. Each subset contains 1001 utterances. To reduce the simulation times, we subsampled the test sets by a factor of 4 (250 utterances per test set, 1000 utterances per SNR). A different subset with 100 utterances from each test set is used for development purposes. All data has a sampling frequency of 8 kHz.

The speech exemplars are extracted from the clean training set. Acoustic feature vectors are represented in the full-resolution STFT domain with K =129 bins and mel-scaled magnitude spectra with 23 frequency bands. The speech exemplars representing half-digits are segmented by a conventional HMM-based system. The recognizer uses in total 53285 speech exemplars distributed to 675 dictionaries of 23 different classes (half-digits plus silence). The number of noise exemplars varies depending on the duration of the noise-only sequences that are selected in the preprocessing step and the estimated SNR level of the target utterance. On average, the recognizer uses 11355 and 6621 noise exemplars/utterance in total at SNR level of -5 dB and 20 dB respectively. The minimum and maximum exemplar lengths are 8 and 40 frames respectively. Exemplars longer than 40 frames are omitted to limit the number of dictionaries. The noise dictionaries are created by performing active noise exemplar selection and noise sniffing [184]. The combined dictionaries and observation matrices are l_2-normalized for all SNR levels. The multiplicative update rule is iterated 100 times for convergence.

The performance of the proposed setup is compared with several baseline speech enhancement systems such as the optimally-modified log-spectral amplitude (OM-LSA) estimator combined with improved minima controlled recursive averaging technique described in [22] and several exemplar-based SR systems described in [12] which use a single overcomplete dictionary containing either fixed length full resolution spectral features (SPEC) or mel-scaled spectral features (MEL). Moreover, the SR-based system adopting the coupled dictionary approach (MELCP) is also considered. The dictionary used by SPEC, MEL and MELCP contains 10000 speech and 10000 noise exemplars. Further details about these systems can be found in [12]. Two evaluation metrics have been used for the performance evaluation. Firstly, the signal-to-distortion ratio (SDR) improvements (ΔSDR) are calculated using the BSS Evaluation Toolkit [173]. Then, the perceptual evaluation of speech quality (PESQ) [139] improvements (ΔPESQ) are also presented to compare the perceptual speech quality of the

proposed system with the baselines.

9.4 Results

The ΔSDR and ΔPESQ values obtained on both test sets of AURORA-2 data are presented in Figure 9.1. Figure 9.1a illustrates the ΔSDR provided on the test set A. The N-REM setup performing the enhancement in mel domain as shown in Equation (9.4) provides ΔSDRs of 10.1 dB, 9.2 dB and 7.9 dB at SNR levels of -5, 0 and 5 dB respectively. The comparable MEL system yields 8.3 dB, 8.4 dB and 7.3 dB at the same SNR levels.

N-REMCP which performs the enhancement in the STFT domain as shown in Equation (9.5), achieves better enhancement than N-REM providing 11.2 dB, 10.2 dB, 8.6dB at SNRs of -5, 0 and 5 dB with an absolute improvement of 1.1 dB, 1.0 dB and 0.7dB. For the same SNRs, the baseline MELCP system provides 10.5 dB, 9.5 dB and 8.1 dB. Both N-REM setups outperform their SR-based counterparts with a considerable margin.

At SNR levels of 10 dB and 15 dB, all systems except OM-LSA provide comparable results with ΔSDRs values between 6.0-6.7 dB at SNR of 10 dB and 4.5-4.8 dB at SNR of 15 dB. The SPEC system outperforms the others with a ΔSDR of 3.3 dB at SNR of 20 dB. OM-LSA provides the worst results at all SNR levels.

The ΔSDR results obtained on test set B, which are shown in Figure 9.1b, clearly demonstrate the improved enhancement provided by N-REM systems especially at lower SNR levels. N-REM provides ΔSDRs of 5.5 dB, 6.8 dB and 6.2 dB at SNRs of -5, 0 and 5 dB. These results are significantly higher than 1.8 dB, 4.8 dB and 4.9 dB of the MEL system. The N-REMCP system outperforms MELCP with an absolute improvement of 2.4 dB, 1.5 dB and 1.2 dB at the same SNRs respectively. N-REM based systems still perform better than the baselines at SNR of 10 dB, while they are slightly worse than MEL and MELCP at 20 dB. At this SNR level, OM-LSA provides the best results with a ΔSDR of 1.6 dB. SPEC is the worst performing system at all SNR levels of test set B.

We further compare the ΔPESQ values to evaluate the perceptual quality of the enhancement systems. The ΔPESQ values obtained on test set A are shown in Figure 9.1c. On test set A at SNR -5 dB, SPEC has the highest ΔPESQ of 0.70 followed by MELCP and N-REMCP with a ΔPESQ of 0.59 and 0.57 respectively. At 0dB, MELCP performs the best with 0.75, while N-REMCP and SPEC yield 0.72 and 0.69 respectively. N-REMCP has the highest ΔPESQ at all SNR levels higher than 0 dB. The performance gap between the N-REM

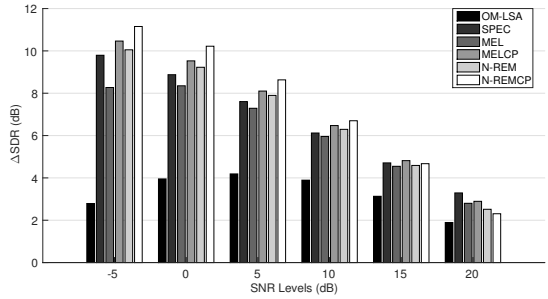
based systems and baselines increases at higher SNR levels. The improved perceptual quality of N-REM and N-REMCP is also apparent from the better ΔPESQ results on test set B at all SNR levels which is shown in Figure 9.1d.

From these results, it can be concluded that the N-REM based systems in general perform better speech enhancement than the baseline systems on account of the separate speech dictionaries which result in more accurate representations of acoustic units in the high-dimensional feature space. Two prominent advantages of these systems are the superior ΔSDR performance under the mismatched noise scenario and overall improvement in the perceptual speech quality. A final comment about the presented results is that the coupled dictionary approach highly improves the enhancement quality also in the N-REM based speech enhancement especially at the lower SNR levels.

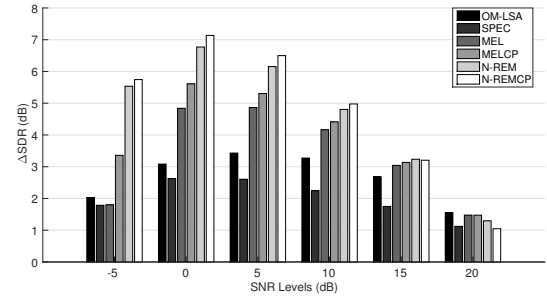
9.5 Conclusion

This chapter presents a novel single-channel speech enhancement system that performs noise robust exemplar matching to separate speech and noise sources using exemplars, each associated with a certain speech unit. These exemplars are organized in separate dictionaries based on the associated speech unit and length and unseen noisy mixtures are approximated as a sparse linear combination of the speech and noise exemplars in each dictionary.

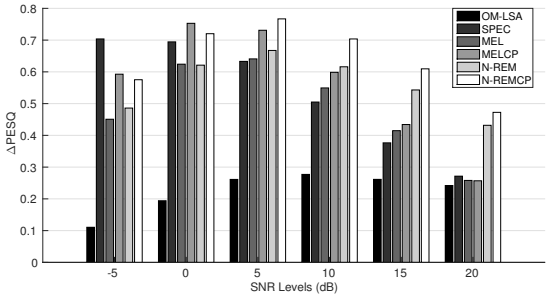
We further adopt the coupled dictionary approach which performs the approximation in the lower dimensional mel domain and the enhancement in the full-resolution STFT domain. The ΔSDR and ΔPESQ results demonstrate the improved speech enhancement achieved by the proposed system.



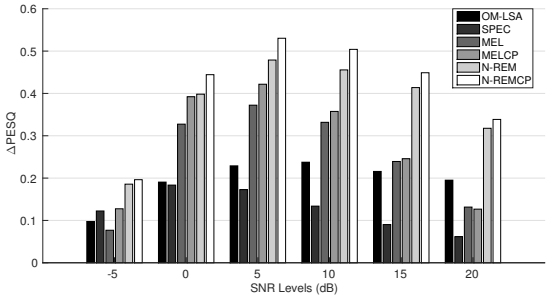
(a) SDR impro. obtained on test set A



(b) SDR impro. obtained on test set B



(c) PESQ impro. obtained on test set A



(d) PESQ impro. obtained on test set B

Figure 9.1: SDR and PESQ improvements on test set A and B of AURORA-2 data

Chapter 10

Applications of N-REM based Speech Enhancement to ASR

We present a novel automatic speech recognition (ASR) scheme which uses the recently proposed noise robust exemplar matching framework for speech enhancement in the front-end. The proposed system employs a GMM-HMM back-end to recognize the enhanced speech signals unlike the prior work focusing on template matching only. Speech enhancement is achieved using multiple dictionaries containing speech exemplars representing a single speech unit and several noise exemplars of the same length. These combined dictionaries are used to approximate the noisy segments and the speech component is obtained as a linear combination of the speech exemplars in the combined dictionaries yielding the minimum total reconstruction error. The performance of the proposed system is evaluated on the small vocabulary track of the 2^{nd} CHiME Challenge and the AURORA-2 database and the results have shown the effectiveness of the proposed approach in improving the noise robustness of a conventional ASR system.

This chapter is adapted from: Emre Yılmaz, Deepak Baby and Hugo Van hamme, "*Noise Robust Exemplar Matching for Speech Enhancement: Applications to Automatic Speech Recognition*", Submitted to INTERSPEECH 2015.

10.1 Introduction

Speech enhancement techniques, aiming to suppress the background noise degrading the speech signals recorded by a microphone, are often combined with automatic speech recognition (ASR) systems for improved noise robustness [100, 107,176]. These techniques reduce the mismatch between the statistical acoustic models, e.g. hidden Markov models (HMM), trained under noise-free conditions and the target speech by preprocessing the noisy speech and/or features to enhance the noise corrupted spectrotemporal structure and recover the speech component as accurately as possible. Numerous enhancement techniques have been combined with Gaussian mixture model (GMM)-HMM [16, 33, 40, 76, 111, 114, 167, 189] and deep neural network (DNN)-HMM [10, 105, 113, 126, 147] ASR systems and reported to provide considerable improvements in the recognition accuracy.

This chapter presents a novel noise robust ASR system which incorporates an exemplar-based speech enhancement approach, dubbed *noise robust exemplar matching* (N-REM), for denoising the target utterance using the actual occurrences of speech and noise extracted from training data. Unlike previous exemplar-based speech enhancement systems using fixed-length exemplars in a single overcomplete dictionary [11, 54, 118, 153], the proposed approach uses exemplars of multiple lengths, each associated with a single speech unit such as phones, syllables, half-words or words [29,36,57]. These exemplars are organized in multiple dictionaries based on the their length and class (associated speech unit). Using separate dictionaries for different speech units is motivated by the geometrical interpretation of SR-based source separation. It is known that the larger the distance between the convex hull of the basis vectors belonging to speech and noise sources are, the better the separation is [37]. Hence, the use of separate dictionaries for each speech unit provides a more precise representation in the high-dimensional feature space.

Previously, the N-REM framework has been successfully applied on small vocabulary ASR tasks [184, 186]. In previous work, the recognizer performs exemplar matching using the mel-scaled spectral representations of the exemplars and noisy speech and relies on a reconstruction error-based back-end to find the most likely hypothesis. However, in the proposed work, N-REM enhances the noisy speech and the enhanced speech represented in the mel frequency cepstral coefficient (MFCC) domain is recognized using a conventional GMM-HMM back-end. This system is expected to remedy the poor recognition accuracy at higher SNR levels thanks to the better discrimination of GMMs trained on MFCC features rather than the suboptimal divergence metric used for exemplar matching. Moreover, on account of the more precise representations of the speech units, the proposed front-end is expected to provide better enhancement and

recognition than the FE approach [50,52] which is an alternative exemplar-based sparse representations approach performing enhancement using fixed-length exemplars in a single overcomplete dictionary. We have performed experiments on both the AURORA-2 database and the small vocabulary track of the 2nd CHiME Challenge to investigate the performance of the proposed approach under different noise and training conditions and compare the performance with other noise robust recognition systems.

10.2 Noise Robust Exemplar Matching

Training frame sequences representing various speech units (speech exemplars) are extracted based on the state-level alignments obtained using an HMM-based recognizer. Speech exemplars, each comprised of D mel frequency bands and spanning l frames, are reshaped into a single vector and stored in the columns of a speech dictionary $\mathbf{S}_{c,l}$: one for each class c and each length l. Each dictionary is of dimensionality $Dl \times N_{c,l}$ where $N_{c,l}$ is the number of available speech exemplars of class c and length l. Similarly, a single noise dictionary $\mathbf{N}_l$ for each length l is formed by reshaping the noise exemplars. Each speech dictionary is concatenated with the noise dictionary of the same length to form a combined dictionary $\mathbf{A}_{c,l} = [\mathbf{S}_{c,l}\, \mathbf{N}_l]$ of dimensionality $Dl \times M_{c,l}$ where $M_{c,l}$ is the total number of available speech and noise exemplars.

Every noisy speech segment of frame length T is also reshaped into vectors by applying a sliding window approach [50] with window length of l frames and stored in an observation matrix $\mathbf{Y}_l = [\mathbf{y}_l^1, \mathbf{y}_l^2 ..., \mathbf{y}_l^{(T-l+1)}]$ of dimensionality $Dl \times (T-l+1)$. Due to multiple-length exemplars, the window length l is varied between the minimum exemplar length $l_{\min}$ and maximum exemplar length $l_{\max}$ yielding observation matrices $\mathbf{Y}_l$ for $l_{\min} \leq l \leq l_{\max}$. For every class c, each observation vector $\mathbf{y}_l$ is expressed as a linear combination of the exemplars that are stored in the dictionaries of the same length: $\mathbf{y}_l \approx \sum_{m=1}^{M_{c,l}} x_{c,l}^m \mathbf{a}_{c,l}^m = \mathbf{A}_{c,l}\mathbf{x}_{c,l}$ for $x_{c,l}^m \geq 0$ where $\mathbf{x}_{c,l}$ is an $M_{c,l}$-dimensional non-negative weight vector. The sparse solutions of $\mathbf{x}_{c,l}$ yield more realistic approximation of the observed segments without overfitting and have been shown to provide better recognition results [80, 174]. The combined dictionaries consisting of speech and noise exemplars are presumed to model all acoustic variability in the observed signal due to pronunciation variation, background noise and so forth. This model can also cope with reverberation by storing reverberated speech exemplars rather than clean speech exemplars.

The non-negative exemplar weights $\mathbf{x}_{c,l}$ are obtained by minimizing the cost function, $d(\mathbf{y}_l, \mathbf{A}_{c,l}\mathbf{x}_{c,l}) + \sum_{m=1}^{M_{c,l}} x_{c,l}^m \Lambda_m$ for $x_{c,l}^m \geq 0$ where $\mathbf{\Lambda}$ is an $M_{c,l}$-

dimensional vector. The first term is the divergence between the observation vector and its approximation. The second term is a regularization term which penalizes the l_1-norm of the weight vector to produce a sparse solution. $\mathbf{\Lambda}$ contains non-negative values and controls how sparse the resulting vector $\mathbf{x}$ is. Defining $\mathbf{\Lambda}$ as a vector, the amount of sparsity enforced on different types of exemplars can be adjusted. In this work, the regularized optimization problem with the aforementioned cost function is solved by applying non-negative sparse coding (NSC) [79]. The generalized KLD is used for d which is commonly used in source separation problems and shown to produce better results than Euclidean distance when used in conjunction with mel-scaled spectral features [174], $d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k$.

All observation matrices $\mathbf{Y}_l$ for $l_{\min} \leq l \leq l_{\max}$ are approximated using the combined dictionaries $\mathbf{A}_{c,l}$ of the same length by applying the multiplicative update rule given in [184]. To quantify the approximation quality, we use the reconstruction error between the noisy speech segments and their approximations. After a fixed number of iterations for all dictionaries, the reconstruction errors between the observation matrix $\mathbf{Y}_l$ and its approximations $\mathbf{A}_{c,l}\mathbf{x}_{c,l}$ are calculated for $l_{\min} \leq l \leq l_{\max}$. As the label of each dictionary is known, decoding is performed by applying dynamic programming to find the class sequence that minimizes the reconstruction error to find the best approximation of the target utterance.

After finding the best approximation, the denoising is performed by reconstructing the frame-wise speech and noise estimates, $\hat{s}_{c,l}$ and $\hat{n}_{c,l}$, that are obtained after removing the windowing effect by adding the components belonging to overlapping windows from the estimates $S_{c,l}X^s_{c,l}$ and $N_l X^n_{c,l}$ respectively. Here, $X^s_{c,l}$ refers to the exemplar weights of the speech exemplars and $X^n_{c,l}$ refers to the exemplar weights of the noise exemplars. The frame-level Wiener-like filter is then obtained as in [11], $W = \mathbf{C}^T \hat{s}_{c,l} \oslash (\mathbf{C}^T(\hat{s}_{c,l} + \hat{n}_{c,l}))$ where $\mathbf{C}$ is the short-time Fourier transform (STFT)-to-mel matrix containing triangular shaped filter-banks.

10.3 Experimental Setup

10.3.1 Databases

AURORA-2: The training material of AURORA-2 [77] database consists of a clean and a multi-condition training set, each containing 8440 utterances. The multi-condition training set was constructed by mixing the clean utterances with noise at SNR levels of 20, 15, 10 and 5 dB. Test set A and B consists of 4

clean and 24 noisy datasets at six SNR levels between -5 and 20 dB. The noise types of test set A match the multi-condition training set. Each subset contains 1001 utterances with one to seven digits 0-9 or oh. We use the complete test sets to be able to compare the results with other systems.
CHIME-2: The small vocabulary track of the 2$^{\text{nd}}$ CHiME Challenge [172] addresses the problem of recognizing commands in a noisy and reverberant living room. The clean utterances contain utterances from 34 speakers reading 6-word sequences of the form *command-color-preposition-letter-digit-adverb*. There are 25 different letters, 10 different digits and 4 different alternatives for each of the other classes. The recognition accuracy of a system is calculated based on the correctly recognized letter and digit keywords.

10.3.2 Dictionary Creation and Implementation Details

AURORA-2: The speech exemplars are extracted from the clean training set. Acoustic feature vectors used during speech enhancement are represented in mel-scaled magnitude spectra with 23 frequency bands. There are in total 52,305 speech exemplars representing half-digits. The minimum and maximum exemplar lengths are 8 and 40 frames respectively. Exemplars longer than 40 frames are omitted to limit the number of dictionaries. The recognizer uses in total 675 dictionaries of 23 different classes (half-digits plus silence). The noise dictionaries are created by performing active noise exemplar selection and noise sniffing which are detailed in [184]. The combined dictionaries and observation matrices are l_3-normalized for all SNR levels. The multiplicative update rule is iterated 100 times for convergence of all frame lengths. Further details can be found in [184].

The enhanced speech is input to a GMM-HMM recognizer employing an HMM topology with 16 states describing each digit and 3 states for silence leading to a total of 179 states. The GMM model is trained on MFCC with 13 static features along with their delta and delta-delta time differences resulting in a 39 dimensional feature space. The emission probabilities of each HMM state is modeled using a GMM of 32 Gaussians with diagonal covariance. For the Viterbi decoder, an HMM topology where all the words have the same word entrance penalties was used. We trained acoustic models on the clean and multicondition training set. To evaluate the impact of retraining on the recognition accuracy, we further train an acoustic model on the enhanced waveforms of the multicondition training set.
CHIME-2: The N-REM system for speech enhancement uses exemplars and noisy speech segments that are represented as mel-scaled magnitude spectral features extracted with a 26 channel mel-scaled filter bank ($D = 26$). The frame length is 25 ms and the frame shift is 10 ms. The binaural data is averaged in

the spectral domain to obtain 26-dimensional feature vectors. The exemplars are extracted from the *reverberated* utterances in the training set according to the state-based segmentations obtained using the acoustic models in the toolkit provided with the database. Exemplars belonging to each speaker are organized in separate dictionary sets for speaker-dependent modeling yielding 34 different dictionary sets. Based on the availability of the exemplars, the minimum and maximum exemplar lengths are 4 and 45 frames respectively.

Half-word exemplars seemed to generalize sufficiently to unseen data. Dictionary sizes vary with different classes and speakers. *Prewarping* [183] is applied to boost the modeling capabilities of the underpopulated speech dictionaries. The number of exemplars in each dictionary after prewarping is limited to 50. The noise dictionaries used for the recognition phase contain 200 noise exemplars that are acquired on the fly from the immediate neighborhood of the target utterance in both directions until the frames belonging to other target utterances. In addition to these sniffed noise exemplars, 200-300 noise exemplars are extracted from the most active 2 noise-only sequences selected by adaptive noise exemplar selection technique [184]. The multiplicative update rule is iterated 25 times to obtain the exemplar weights. The columns of the combined dictionaries and observation matrices are l_2-normalized. Further details can be found in [184].

The enhanced speech is recognized using the baseline HMM structure provided by the challenge organizers at the back-end [172]. The provided acoustic models use 4-10 state word-level HMMs and the emission probabilities of each HMM state is modeled using a GMM of 7 Gaussians. The speech features are standard 39-dimensional MFCCs applied with cepstral mean normalization. We first use the default acoustic models trained on clean, reverberated and noisy training utterances. Similar to the AURORA-2 experiments, we also retrain a new acoustic model using the enhanced isolated noisy training utterances.

10.3.3 Evaluation Metrics

We have opted for the metrics which have been traditionally used for the evaluation of the databases described in Section 10.3.1 for comparability with the previous literature. The word error rate (WER) has been used to quantify the recognition accuracy for the AURORA-2 digit recognition task. The keyword recognition accuracy (RA) is used to evaluate the system performance on the CHIME-2 data.

Table 10.1: Word error rates in % obtained on test set A and B of AURORA-2 data

(a) Test set A

SNR(dB)	-5	0	5	10	15	20	*0-20*	clean
N-REM	20.1	10.0	6.3	4.6	3.5	2.7	5.4	1.8
N-REM-SE (clean)	24.4	10.1	5.3	3.4	2.1	1.4	4.4	**0.3**
N-REM-SE (multi)	**17.9**	**8.2**	**4.5**	**2.8**	**1.9**	**1.0**	**3.6**	0.8
N-REM-SE (retrain)	19.5	8.6	4.7	3.2	**1.9**	**1.0**	3.9	0.5

(b) Test set B

SNR(dB)	-5	0	5	10	15	20	*0-20*	clean
N-REM	56.9	25.6	10.4	5.7	3.8	3.2	9.7	1.8
N-REM-SE (clean)	56.3	24.1	9.5	4.2	2.3	1.3	8.3	**0.3**
N-REM-SE (multi)	**55.0**	**23.9**	**8.9**	**3.9**	**1.9**	**1.2**	**8.0**	0.8
N-REM-SE (retrain)	55.7	24.1	9.3	4.1	2.1	**1.2**	8.2	0.5

10.4 Results

We perform recognition experiments using the proposed system (N-REM-SE) on test set A and B of AURORA-2 and development and test sets of CHIME-2 data. For both datasets, we first compare the performance of N-REM-SE with the exemplar matching version (N-REM) [184] using the same combined dictionaries and divergence measure. Then, the proposed system is compared with other noise robust ASR systems to evaluate the overall performance of N-REM-SE.

10.4.1 AURORA-2 Results

Table 10.1 and 10.2 presents the results obtained on AURORA-2 data. The clean speech recognition performance of all systems is given in the last column of Table 10.1a and Table 10.1b. The exemplar matching system provides a WER of 1.8% on clean speech which is higher than any setup with a GMM-HMM back-end. N-REM-SE trained on clean and multicondition data has a WER of 0.3 and 0.8 on clean speech respectively. As expected, the GMM-HMM back-end considerably improves the clean speech recognition performance.

N-REM provides WERs of 20.1%, 10.0% and 6.3% at SNR levels of -5 dB , 0 dB and 5 dB. N-REM-SE trained on clean speech performs surprisingly well with WERs of 24.4%, 10.1% and 5.3% at the same SNR levels. Training the

Table 10.2: Comparison of NREM-SE with other recognition systems on AURORA-2 data

Technique	*test set A*		*test set B*	
	-5	*0-20*	-5	*0-20*
GMM-HMM [52]	77.2	16.9	77.1	15.9
AFE [78]	56.5	7.7	57.7	8.2
NAT [90]	57.7	6.3	58.1	6.3
SC [52]	35.7	7.2	49.8	9.3
FE [52]	30.4	3.6	50.8	6.1
SC+FE [52]	25.6	3.1	**43.7**	5.0
ESSEM-MCM [166]	-	4.4	-	**4.7**
RBM-DNN [106]	-	4.5	-	5.1
MASK-RBM-DNN [105]	-	3.8	-	5.0
MS-CD [12]	21.1	**2.4**	62.4	7.5
FE+MS-CD [12]	20.6	**2.4**	54.2	6.1
N-REM [184]	20.1	5.4	56.9	9.7
N-REM-SE	**17.9**	3.6	55.0	8.0

acoustic models of N-REM-SE on multicondition data improves the results by an absolute improvements of 6.5%, 1.9% and 1.0% respectively. Unlike the other exemplar-based approaches which use a single fixed noise dictionary, retraining the acoustic models on the enhanced training data does not bring any improvement in case of N-REM-SE. This is due to the adaptive noise modeling adopted in N-REM-SE which selects a different set of noise exemplars for each noisy utterance on the fly. Consequently, retraining does not help the back-end to cope with the artifacts introduced by speech enhancement in this scenario. The models trained on multicondition data yield better or similar recognition performance at all SNR levels.

Using a GMM-HMM back-end reduces the WER in general at higher SNR levels similar to the clean speech performance. Compared to the WERs of 4.6%, 3.5% and 2.7% provided by N-REM at SNRs of 10 dB, 15 dB and 20 dB respectively, N-REM-SE trained on multicondition data has WERs of 2.8%, 1.9% and 1.0% at the same SNRs. Moreover, this system has an average WER (*0-20*) of 3.6% compared to the 5.4% of N-REM. Multicondition trained N-REM-SE also shows superior performance at all SNR levels of test set B compared to N-REM and other N-REM-SE variants. This system provides an average WER of 8.0% compared to the 9.7% of N-REM and 8.3% of retrained N-REM-SE.

Table 10.2 lists the recognition results of some other noise robust ASR systems data with state-of-the-art performance on AURORA-2 data. This list is by

no means exhaustive and it only includes the recognition results published on the complete test sets for a fair comparison. From these results, it can be concluded that the recognition systems using exemplar-based speech enhancement approaches, e.g. FE variants and N-REM-SE, provide impressive performance in matched noise scenarios. Other exemplar-based systems which do not rely on a statistical model at the back-end, e.g. SC and N-REM, mainly suffer from low recognition accuracies at higher SNR levels resulting in worse average WER results. On the other hand, the ESSEM-MCM and RBM-DNN methods perform almost equally well under matched and mismatched noise conditions.

The hybrid SC+FE system appears to be a nice compromise with a remarkable -5 dB performance and one of the lowest average WERs on both test sets. Compared to this system and other exemplar-based systems, the gap between the matched and mismatched noise is larger for the proposed system due to the smaller amount of noise exemplars in the class- and length-dependent dictionaries. This results in poor generalization against unseen noise types. In the case of matched noise, N-REM-SE has a better -5 dB performance, which is actually the best among all systems, and a comparable average WER on test set A.

10.4.2 CHIME-2 Results

Table 10.3 and 10.4 presents the results obtained on CHIME-2 data. We first focus on Table 10.3 presenting the recognition accuracies obtained on the development and test sets to compare the performance of the exemplar matching-based system and the proposed recognizer. The results on both sets follow a similar trend; hence the results on the test set are discussed only. N-REM provides RAs of 71.0%, 78.9% and 85.3% at -6 dB, -3 dB and 0 dB. N-REM-SE trained on reverberated data has RAs of 69.8%, 76.8% and 84.3% at the same SNR levels. These results are slightly worse than the exemplar matching system. Retraining the acoustic models does not improve the recognition performance similar to the results obtained on AURORA-2 data. The proposed setup with the acoustic models trained on clean and noisy speech provides inferior results.

The overall performance of N-REM is also mildly better with an average RA (Avg) of 84.8% compared to the proposed recognizer trained on the reverberated data with 83.4%. The CHIME-2 results favor the exemplar matching system over N-REM-SE unlike the AURORA-2 experiments. The same observation holds for the single dictionary counterparts, FE and SC [50], considering the recognition results reported in [184]. We discuss the differences between AURORA-2 and CHIME-2 databases to get more insight for the reduced performance of the

Table 10.3: Keyword recognition accuracies in % obtained on the development and test set of CHIME-2 data

(a) Development Set

SNR(dB)	-6	-3	0	3	6	9	*Avg*
N-REM	**70.4**	**77.9**	**84.8**	**90.4**	**92.6**	**93.8**	**85.0**
N-REM-SE (clean)	21.1	23.4	27.9	30.5	34.4	34.9	28.7
N-REM-SE (reverb)	67.3	74.7	81.7	**88.2**	**89.8**	91.5	82.2
N-REM-SE (noisy)	60.8	68.8	74.3	78.1	81.8	83.2	74.5
N-REM-SE (retrain)	**69.4**	**75.9**	**82.8**	87.4	88.6	**91.7**	**82.6**

(b) Test set

SNR(dB)	-6	-3	0	3	6	9	*Avg*
N-REM	**71.0**	**78.9**	**85.3**	**88.7**	**91.9**	**92.8**	**84.8**
N-REM-SE (clean)	19.9	22.8	26.8	29.7	34.2	38.1	28.6
N-REM-SE (reverb)	69.8	**76.8**	84.3	**87.3**	**90.3**	**91.6**	**83.4**
N-REM-SE (noisy)	60.3	69.1	74.8	78.0	81.4	82.8	74.4
N-REM-SE (retrain)	**70.3**	76.4	**84.5**	86.7	89.3	90.6	83.0

Table 10.4: Comparison of NREM-SE with other recognition systems on CHIME-2 data

Technique	*Baseline AM*		*Retrained AM*	
	Dev Avg	Test Avg	Dev Avg	Test Avg
GMM [172]	68.7	68.8	-	-
SCSS [123]	76.0	77.7	-	-
FE [51]	81.2	82.2	-	-
HMM-FE [51]	81.9	82.7	-	-
BSE [127]	81.8	82.0	81.5	83.2
FASST [164]	**82.9**	**84.2**	**84.7**	**85.7**
N-REM-SE	82.2	83.4	82.6	83.0

exemplar-based front-end denoising systems (N-REM-SE and FE) on CHIME-2 compared to the other systems (N-REM and SC) which do not rely on a statistical model at the back-end. Firstly, the variations in both speech and noise components in the noisy mixtures are more significant in CHIME-2 compared to AURORA-2. The former is due to the smearing effect of reverberation degrading the spectrotemporal content of the speech exemplars and the latter is an outcome of the highly non-stationary room noise. Secondly, there are

only few exemplars available in the training data, especially for *letters*, to obtain accurate representations of each speech unit in the high-dimensional feature space. Hence, the speech enhancement quality provided by the combined dictionaries with increased variation is less effective in compensating for the mismatch between the target speech and the acoustic models trained on neither reverberated nor noisy speech. Under such a scenario, adopting a GMM-HMM back-end is less favorable compared to the N-REM and SC systems which either rely on the reconstruction error or estimate state likelihoods directly from the exemplar weights at the back end respectively.

To be able to evaluate the enhancement performance of the N-REM front-end, we present some results obtained using other speech or feature enhancement-based recognition systems in Table 10.4. The recognition results of the best performing GMM-HMM baseline trained reverberated data is also provided as a reference. The best performing FASST system does not only benefit from spectral enhancement, but also from spatial enhancement using spatial full-rank covariance matrices [164]. All other systems benefiting only from either spectral or spatial enhancement and using the standard acoustic models provided as a part of the CHIME-2 challenge perform moderately compared to the more sophisticated approaches with some speaker and environment adaptation techniques. N-REM-SE provides a reasonable performance outperforming the other exemplar-based sparse representation systems, FE and HMM-FE, which use exemplars of fixed length in a single overcomplete dictionary for feature enhancement.

10.5 Conclusion

This chapter presents a novel noise robust ASR system using a single-channel speech enhancement setup that performs noise robust exemplar matching to separate speech and noise sources in the front-end. The exemplars used in this technique are associated with a certain speech unit and organized in separate dictionaries based on the associated speech unit and length. The noisy mixtures are approximated as a sparse linear combination of the speech and noise exemplars in each dictionary. The proposed system has provided comparable performance with the other state-of-the-art ASR systems on two popular small vocabulary recognition tasks, AURORA-2 and CHIME-2.

Chapter 11

Conclusion

This chapter concludes the thesis by providing a review of the original contributions and several directions for future research.

11.1 Original Contributions

This work focuses on establishing a noise modeling scheme for traditional exemplar matching-based automatic speech recognition (ASR) systems. This novel noise robust ASR framework is the main contribution of this thesis. Furthermore, the attempts to improve the speech and noise modeling capabilities and reduce the computational burden yield novel data selection techniques that can also be used in other applications. The list of all original contributions is given below.

- **Noise robust exemplar matching (N-REM) framework**: The idea of organizing exemplars of different lengths in multiple dictionaries and approximate noisy speech as a linear combination of these exemplars for noise robust automatic speech recognition is the main contribution of this thesis to the ASR literature. The initial efforts are explained in Chapter 2 and the first complete system is the topic of Chapter 5. The investigation of the alpha-beta divergence in place of the generalized Kullback-Leibler divergence is another important novelty bringing improved noise robustness. The details of this work are given in Chapter 7.

- **Time warping for N-REM**: The traditional exemplar matching greatly benefits from dynamic time warping (DTW) as DTW allows the comparison of segments of different frame length. Motivated by this fact, we have also proposed a time warping technique for the N-REM framework and investigated the influence on clean speech recognition in Chapter 3. Despite the promising results, the proposed time warping technique is not adopted in the latter steps due to large computational complexity.
- **Data selection for N-REM**: Data selection from the N-REM dictionaries is also addressed in this thesis. The exemplar selection experiments aim to find out the optimal acoustic model size in terms of the number of speech exemplars used for recognition. Several exemplar selection criteria for the N-REM dictionaries are introduced in Chapter 4 and 8. The former techniques use the generalized Kullback-Leibler divergence as a dissimilarity measure, while the latter adopts the alpha-beta divergence in accordance with the recognizer.
- **Adaptive noise modeling for N-REM**: The recognition performance of the proposed technique heavily depends on the use of noise exemplars that can accurately capture the characteristics of the actual noise signal corrupting the target speech. For this reason, the noise exemplars are extracted on the fly based on the information obtained in a preprocessing stage and a unique noise dictionary is used for each utterance. The details of this adaptive noise modeling approach is given in Chapter 6.
- **N-REM-based Speech Enhancement**: The novel recognition architecture is applied for denoising the speech signals and the results are presented in terms of the traditional speech enhancement measures in Chapter 9. Finally, this enhancement system is integrated into a conventional GMM-HMM speech recognizer to reduce the amount of the background noise in the front-end. The recognition results are presented in Chapter 10.

11.2 Directions for Future Research

This section discusses the possible extensions to the described framework for more widespread use, improved recognition performance and reduced computational complexity.

- **Hybrid acoustic modeling combined with statistical models**: The performance of the N-REM system, which only relies on the reconstruction error at the back-end, can be improved by adopting other statistical

acoustic models in parallel. In such a setting, the acoustic scores obtained from both streams can be combined to benefit from the noise robustness of exemplar-based acoustic modeling and better discrimination of the statistical models such as complex GMM distributions in conjunction with MFCC features or DNNs. One possible future direction is to develop a robust way of combining the acoustic scores that are obtained using different models, such as the reconstruction errors of exemplar matching and the likelihoods of the GMMs or "pseudo-likelihoods" of the DNNs. This joint approach is shown to be effective, e.g. in [1, 29, 55].

- **Efficient algorithms to obtain the exemplar weights**: The computational bottleneck of N-REM is the multiplicative update rule that is applied iteratively to learn the exemplar weights. There are several algorithms for computationally efficient estimation of the exemplar weights such as [58, 95, 169, 175]. Adopting such an efficient exemplar weight learning technique will reduce the computational burden and make it viable to model a larger number of acoustic units, such as context-dependent phones rather than half-words.

- **Extention to large vocabulary tasks**: The presented work is a first step toward a noise robust exemplar matching framework which can handle medium-large vocabulary speech. Its main advantage is that the exemplars model speech units, which should scale better than the long, fixed-length exemplars employed in other exemplar-based sparse representations systems. In this thesis, the speech exemplars have been chosen to represent half-words. Considering the dimensionality and computational restrictions, the same framework using exemplars associated with more general subword units such as phones, syllables could be applied to a medium- or large-vocabulary task. Only the current decoding scheme would need to be redesigned in a way that it will incorporate a language model combined with the acoustic costs, but for this it could largely rely on existing exemplar matching frameworks [30].

- **Further investigation of the time warping**: Even though the feasibility of the proposed time warping model has been shown, there are still more open questions such as the different warping matrix designs and their effects on the recognition accuracy, a detailed analysis of the effect of different sparsity factors on the recognition accuracy, tying the weights of the time-frequency cells belonging to the same frame to obtain a frame-level time warping and designing a dedicated implementation of the proposed model which is expected to reduce the simulation times. Furthermore, the proposed time warping technique is not applied under noisy scenarios and its impact on the recognition accuracy is still unknown. This aspect has to be further investigated after tackling the computational

restrictions due to the increased number of exemplar weights that has to be learned.

- **Automatic discovery of the divergence parameters**: Another future work is to develop a procedure to yield the optimal parameters of the alpha-beta divergence in terms of the recognition performance. As reported in Chapter 7, the N-REM recognizer using the alpha-beta divergence provided considerable improvements in the recognition accuracy especially at the lower SNR levels. However, the divergence parameters (α, β) are tuned manually on development data which requires a time-consuming search over the (α, β) plane. The proposed recognizer will also benefit from estimating the divergence parameters without any prior investigation, as the divergence parameters providing the best performance depend highly on the characteristics of noisy mixtures [20]. An asymmetric clustering algorithm using the AB divergence has been recently proposed which estimates the divergence parameters based on the within-cluster variances [130]. A systematic and computationally efficient way of estimating the divergence parameters from the training data would reduce the computational burden due to the initial search for suitable values.

Appendix A

Appendix A

A.1 Silence Compensation

Silence compensation is performed by applying non-negative sparse coding using a single dictionary $\mathbf{A}^*_{L_s}$ containing speech exemplars from all classes and noise exemplars from different noise-only training sequences as illustrated in Figure 5.2.

$$\mathbf{Y}_{L_s} \approx \mathbf{A}^*_{L_s}\mathbf{x}_{L_s} \qquad \text{s.t.} \qquad \mathbf{x}_{L_s} \geq 0. \tag{A.1}$$

The SNR level is estimated as the ratio of total speech weights and total speech and noise weights in order to limit the estimation range to [0,1]

$$\text{SNR}_{\text{est}} = \frac{\sum_{w=1}^{W}\sum_{m=1}^{J} x_{L_s}^{w,m}}{\sum_{w=1}^{W}\sum_{m=1}^{M} x_{L_s}^{w,m}} \cdot F. \tag{A.2}$$

$\mathbf{x}^w_{L_s}$ is the sparse weight vector corresponding to w^{th} of W noisy segments of length L_s. J is the number of speech exemplars and M is number of all exemplars excluding the sniffed noise exemplars. In order to balance the bias in case of mismatched noise, SNR_{est} value is scaled by a constant, F, which is calculated as the ratio of the total weight obtained by the most active 10% of the noise-only training sequences to the total weights. In case of matched noise, F is very close to 1 as the actual noise component is approximated as a superposition of noise exemplars from a few training sequences only. On the other hand, approximation of the mismatched noise requires a larger number of training sequences resulting in a smaller F.

After obtaining the weights by applying the multiplicative update rule in Equation (5.4), the reconstruction of the speech is obtained by linearly combining the speech exemplars only. A frame-level estimation of the speech activity (FSA) is obtained by summing over the frequency bins of the reconstructed speech, normalizing the values to the [0,1] range over the complete utterance and inverting such that 1 denotes the silence and 0 denotes the maximum observed speech activity. Then, in order to obtain steeper transitions between speech and silence regions, we calculate the speech activity value (VAD) by applying a shifted and scaled logistic function [50] to the FSA values

$$\mathrm{VAD} = \frac{1}{1 + \exp(c_1 \cdot \mathrm{FSA} - \beta)} \tag{A.3}$$

with the parameters c_1 and β. c_1 is a scalar and set to 10. β is an SNR-dependent value which is calculated as

$$\beta = \frac{1 - \exp(c_1 \cdot \zeta)}{\exp(c_1 \cdot \zeta) - \exp(c_1)} \tag{A.4}$$

$$\zeta = \min(\mathrm{SNR}_{\mathrm{est}} \cdot c_2, c_3)) \tag{A.5}$$

where c_2 and c_3 are set to 1.5 and 0.55 respectively. Thresholding is applied to the VAD values to obtain a binary decision for each frame

$$\mathrm{VAD}_{\mathrm{h}} = \begin{cases} 1 & \text{if VAD}{>}\mathrm{VAD}_{\mathrm{thr}} \\ 0 & \text{if VAD}{\leq}\mathrm{VAD}_{\mathrm{thr}}. \end{cases}$$

$\mathrm{VAD}_{\mathrm{thr}}$ is set to 0.95. The reconstruction errors corresponding to the silence dictionaries are scaled by a value CF which depends on the $\mathrm{VAD}_{\mathrm{h}}$ value assigned to the middle frame of the corresponding noisy segment and the SNR estimate,

$$\mathrm{CF} = 1 - \min(\max(\mathrm{SNR}_{\mathrm{est}} \cdot \theta, \phi), \gamma) \cdot \mathrm{VAD}_{\mathrm{h}} \tag{A.6}$$

where θ is a scale factor, ϕ and γ are lower and upper limits. They are set to 0.75, 0.1 and 0.55 respectively.

A.2 SNR-dependent ANES

The single dictionary, which is illustrated in Figure 5.2, contains noise exemplars extracted from 800 noise-only training sequences (10 exemplars from each sequence). Once the most active training sequences are found, i.e. the noise sequences obtaining the highest weights, noise dictionaries that are used in the recognition phase are extracted from the N_{max} most active training sequences

with hops of 3 frames (between 77-170 exemplars from each sequence). The value of $N_{\max}$ depends on the SNR level and it is chosen according to the $\mathrm{SNR}_{\mathrm{est}}$ parameter (cf. Appendix A.1):

$$N_{\max} = \begin{cases} 3 & \text{if } \mathrm{SNR}_{\mathrm{est}} < 0.35 \\ 2 & \text{if } \mathrm{SNR}_{\mathrm{est}} \geq 0.35 \text{ and } \mathrm{SNR}_{\mathrm{est}} < 0.5 \\ 1 & \text{if } \mathrm{SNR}_{\mathrm{est}} \geq 0.5 \text{ and } \mathrm{SNR}_{\mathrm{est}} < 0.65 \\ 0 & \text{if } \mathrm{SNR}_{\mathrm{est}} \geq 0.65. \end{cases}$$

The combined dictionaries of all classes and lengths contain the noise exemplars that are extracted from these $N_{\max}$ training sequences. The combined dictionaries contain only a few or no noise exemplars during the recognition of high SNR levels.

Bibliography

[1] Aradilla, G., Vepa, J., and Bourlard, H. Improving speech recognition using a data-driven approach. In *Proc. INTERSPEECH* (Lisbon, Portugal, 2005), pp. 3333–3336. pages 24, 55, 90, 149

[2] Aradilla, G., Vepa, J., and Bourlard, H. Using posterior-based features in template matching for speech recognition. In *International Conference on Spoken Language Processing (ICSLP)* (2006), pp. 2570–2573. pages 10, 34

[3] Arai, T., and Greenberg, S. Speech intelligibility in the presence of cross-channel spectral asynchrony. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 1998), pp. 933 –936. pages 35

[4] Ari, I., Cemgil, A. T., and Akarun, L. Probabilistic interpolative decomposition. In *IEEE Workshop on Machine Learning for Signal Processing (MLSP)* (Santander, Spain, Sept. 2012), pp. 1–6. pages 45

[5] Ari, I., Simsekli, U., Cemgil, A. T., and Akarun, L. Randomized matrix decompositions and exemplar selection in large dictionaries for polyphonic piano transcription. *Journal of New Music Research 43(3)* (2014), 255–265. pages 115

[6] Astudillo, R. F., Hoffmann, E., Mandelartz, P., and Orglmeister, R. Speech enhancement for automatic speech recognition using complex gaussian mixture priors for noise and speech. In *NOLISP* (2009), J. S. i Casals and V. Zaiats, Eds., vol. 5933 of *Lecture Notes in Computer Science*, Springer, pp. 60–67. pages 16

[7] Atal, B. S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America 55*, 6 (1974), 1304–1312. pages 3

[8] Austin, S., Schwartz, R., and Placeway, P. The forward-backward search algorithm. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Apr. 1991), pp. 697–700 vol. 1. pages 13

[9] Axelrod, S., and Maison, B. Combination of hidden Markov models with dynamic time warping for speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2004), vol. 1, pp. 173–176. pages 24, 55

[10] Baby, D., Gemmeke, J. F., Virtanen, T., and Van hamme, H. Exemplar-based speech enhancement for deep neural network based automatic speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2015). pages 136

[11] Baby, D., Virtanen, T., Barker, T., and Van hamme, H. Coupled dictionary training for exemplar-based speech enhancement. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2014), pp. 2883–2887. pages 136, 138

[12] Baby, D., Virtanen, T., Gemmeke, J. F., Barker, T., and Van hamme, H. Exemplar-based noise robust automatic speech recognition using modulation spectrogram features. In *Proc. IEEE Spoken Language Technology Workshop* (South Lake Tahoe, USA, Dec. 2014). pages 15, 126, 129, 130, 142

[13] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. Automatic speech recognition and speech variability: A review. *Speech Communication 49*, 10–11 (2007), 763–786. pages 90

[14] Berndt, D., and Clifford, J. Using dynamic time warping to find patterns in time series. In *Workshop on Knowledge Discovery in Databases (KDD)* (1994), pp. 359–370. pages 34

[15] Bourlard, H., Hermansky, H., and Morgan, N. Towards increasing speech recognition error rates. *Speech Communication 18*, 3 (1996), 205 – 231. pages 90

[16] Breithaupt, C., and Martin, R. Statistical analysis and performance of DFT domain noise reduction filters for robust speech recognition. In *Proc. INTERSPEECH* (2006), pp. 365–368. pages 136

[17] Carmi, A., Gurfil, P., Kanevsky, D., and Ramabhadran, B. ABCS: Approximate Bayesian Compressed Sensing. Tech. rep., Human Language Technologies, IBM, 2009. pages 59

[18] Chen, S. F., and Goodman, J. An Empirical Study of Smoothing Techniques for Language Modeling, 1998. pages 13

[19] Christensen, H., Barker, J., Ma, N., and Green, P. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In *Proc. INTERSPEECH* (Makuhari, Japan, Sept. 2010). pages 109

[20] Cichocki, A., Cruces, S., and Amari, S.-I. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy 13*, 1 (2011), 134–170. pages 91, 94, 95, 112, 115, 150

[21] Cichocki, A., Zdunek, R., and Amari, S. Csiszár's divergences for non-negative matrix factorization: family of new algorithms. In *Proc. of the 6th International Conference on Independent Component Analysis and Blind Signal Separation* (2006), pp. 32–39. pages 60, 95

[22] Cohen, I. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing 11*, 5 (Sept 2003), 466–475. pages 130

[23] Cooke, M., Barker, J., Cunningham, S., and Shao, X. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 5 (2006), 2421–2424. pages 65, 99

[24] Cooke, M., Green, P., Josifovski, L., and Vizinho, A. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication 34*, 3 (2001), 267–285. pages 16, 111

[25] Cooke, M., Morris, A., and Green, P. Missing data techniques for robust speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Apr 1997), vol. 2, pp. 863–866 vol.2. pages 16

[26] Dahl, G., Yu, D., Deng, L., and Acero, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 20*, 1 (Jan. 2012), 30–42. pages 7, 9

[27] Davis, S., and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing 28*, 4 (Aug. 1980), 357–366. pages 3

[28] De Wachter, M., Demuynck, K., and Van Compernolle, D. Boosting HMM performance with a memory upgrade. In *International Conference on Spoken Language Processing (ICSLP)* (2006), pp. 1730–1733. pages 55

[29] De Wachter, M., Demuynck, K., Van Compernolle, D., and Wambacq, P. Data driven exemplar based continuous speech recognition. In *Proc. European Conf. on Speech Communication and Technology* (2003), pp. 1133–1136. pages 10, 34, 55, 80, 90, 114, 126, 136, 149

[30] De Wachter, M., Matton, M., Demuynck, K., Wambacq, P., Cools, R., and Van Compernolle, D. Template-based continuous speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 4 (May 2007), 1377–1390. pages 24, 25, 26, 34, 44, 55, 63, 90, 149

[31] Demuynck, K. *Extracting, Modelling and Combining Information in Speech Recognition.* PhD thesis, K.U.Leuven, ESAT, February 2001. pages 3, 24

[32] Deng, L., Acero, A., Jiang, L., Droppo, J., and Huang, X. High-performance robust speech recognition using stereo training data. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2001), vol. 1, pp. 301–304. pages 16

[33] Deng, L., Acero, A., Plumpe, M., and Huang, X. Large-vocabulary speech recognition under adverse acoustic environments. In *Proc. ICSLP* (Oct. 2000), pp. 806–809. pages 136

[34] Deng, L., and Strik, H. Structure-based and template-based automatic speech recognition - comparing parametric and non-parametric approaches. In *Proc. INTERSPEECH* (Aug. 2007), pp. 898–901. pages 55

[35] Deng, W., Qian, Y., Fan, Y., Fu, T., and Yu, K. Stochastic data sweeping for fast DNN training. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2014), pp. 240–244. pages 9

[36] Deselaers, T., Heigold, G., and Ney, H. Speech recognition with state-based nearest neighbour classifiers. In *Proc. INTERSPEECH* (Antwerp, Belgium, 2007), pp. 2093–2096. pages 55, 90, 114, 126, 136

[37] Donoho, D., and Stodden, V. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004. pages 57, 92, 126, 136

[38] Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Relative-error CUR matrix decompositions. *SIAM J. Matrix Anal. Appl. 30*, 2 (Sept. 2008), 844–881. pages 45, 114

[39] Droppo, J., Acero, A., and Deng, L. Uncertainty decoding with SPLICE for noise robust speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2002), pp. 57–60. pages 16

[40] ETSI. *ETSI ES 202 050 V1.1.5 (2007-01), Advanced front-end feature extraction algorithm*, January 2007. pages 136

[41] Fazel, A., and Chakrabartty, S. Sparse auditory reproducing kernel (spark) features for noise-robust speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on 20*, 4 (May 2012), 1362–1371. pages 15

[42] Févotte, C., Bertin, N., and Durrieu, J.-L. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Comput. 21*, 3 (Mar. 2009), 793–830. pages 110

[43] Forney, G.D., J. The Viterbi algorithm. *Proceedings of the IEEE 61*, 3 (March 1973), 268–278. pages 13

[44] Frieze, A., Kannan, R., and Vempala, S. Fast Monte-Carlo algorithms for finding low-rank approximations. *J. ACM 51*, 6 (Nov 2004), 1025–1041. pages 45, 114

[45] Gales, M. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language 12*, 2 (1998), 75 – 98. pages 16

[46] Gales, M. J. F., and Young, S. J. Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing 4*, 5 (Sept. 1996), 352–359. pages 16, 17

[47] Gauvain, J., and Lee, C.-H. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing 2*, 2 (Apr 1994), 291–298. pages 16

[48] Gemmeke, J., ten Bosch, L., Boves, L., and Cranen, B. Using sparse representations for exemplar based continuous digit recognition. In *Proc. EUSIPCO* (Glasgow, Scotland, August 24–28 2009), pp. 1755–1759. pages 24, 26, 34, 44, 56

[49] Gemmeke, J., and Virtanen, T. Noise robust exemplar-based connected digit recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (March 2010), pp. 4546–4549. pages 24, 44, 56

[50] Gemmeke, J., Virtanen, T., and Hurmalainen, A. Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 19*, 7 (Sept. 2011), 2067–2080. pages 11, 27, 29, 34, 36, 38, 46, 57, 60, 61, 71, 72, 80, 81, 90, 92, 96, 104, 114, 116, 120, 126, 127, 137, 143, 152

[51] Gemmeke, J. F., Hurmalainen, A., and Virtanen, T. HMM-regularization for NMF-based noise robust ASR. In *2nd International Workshop on Machine Listening in Multisource Environments* (2013), pp. 47–52. pages 71, 107, 144

[52] Gemmeke, J. F., and Van hamme, H. Advances in noise robust digit recognition using hybrid exemplar-based techniques. In *Proc. INTERSPEECH* (Portland, USA, Sept. 2012), pp. 1–4. pages 64, 71, 82, 84, 98, 137, 142

[53] Gemmeke, J. F., and Virtanen, T. Artificial and online acquired noise dictionaries for noise robust ASR. In *Proc. INTERSPEECH* (2010), pp. 2082–2085. pages 63, 64, 82, 98

[54] Gemmeke, J. F., Virtanen, T., and Hurmaleinen, A. Exemplar-based speech enhancement and its application to noise-robust automatic speech recognition. In *International Workshop on Machine Listening in Multisource Environments* (Sept. 2011), pp. 53–75. pages 16, 34, 44, 56, 90, 136

[55] Ghitza, O., and M., S. M. Hidden Markov models with templates as non-stationary states: an application to speech recognition. *Computer Speech & Language 7*, 2 (1993), 101–119. pages 55, 149

[56] Glasbey, C. A., and Mardia, K. V. A review of image-warping methods. *Journal of Applied Statistics 25*, 2 (1998), 155–171. pages 34

[57] Golipour, L., and O'Shaughnessy, D. Context-independent phoneme recognition using a k-nearest neighbour classification approach. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Apr. 2009), pp. 1341–1344. pages 24, 25, 55, 90, 114, 126, 136

[58] Gong, P., and Zhang, C. Efficient nonnegative matrix factorization via projected Newton method. *Pattern Recognition 45*, 9 (2012), 3557 – 3565. pages 149

[59] Gong, Y. Speech recognition in noisy environments: A survey. *Speech Communication 16*, 3 (1995), 261–291. pages 24

[60] Gong, Y. A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition. *IEEE Transactions on Speech and Audio Processing 13*, 5 (Sept 2005), 975–983. pages 16

[61] Goreinov, S. A., Tyrtyshnikov, E. E., and Zamarashkin, N. L. A theory of pseudoskeleton approximations. *Linear Algebra and its Applications 261*, 1–3 (1997), 1–21. pages 45, 114

[62] Grancharov, V., Samuelsson, J., and Kleijn, B. On causal algorithms for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing 14*, 3 (2006), 764–773. pages 126

[63] Hakkani-Tur, D., Riccardi, G., and Gorin, A. Active learning for automatic speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2002), vol. 4, pp. IV–3904–IV–3907. pages 114

[64] Halko, N., Martinsson, P.-G., and Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review 53*, 2 (2011), 217–288. pages 44, 63, 114

[65] Han, J., Kamber, M., and Tung, A. K. H. Spatial clustering methods in data mining: A survey. In *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS* (2001). pages 118

[66] Hannemann, M., Povey, D., and Zweig, G. Combining forward and backward search in decoding. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2013), pp. 6739–6743. pages 13

[67] Hanson, B., and Applebaum, T. Subband or cepstral domain filtering for recognition of lombard and channel-distorted speech. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (April 1993), vol. 2, pp. 79–82. pages 15

[68] Hariharan, R., Kiss, I., and Viikki, O. Noise robust speech parameterization using multiresolution feature extraction. *IEEE Transactions on Speech and Audio Processing 9*, 8 (Nov 2001), 856–865. pages 15

[69] Hecht-Nielsen, R. Theory of the backpropagation neural network. In *Neural Networks, 1989. IJCNN., International Joint Conference on* (1989), pp. 593–605 vol.1. pages 9

[70] Heigold, G., Nguyen, P., Weintraub, M., and Vanhoucke, V. Investigations on exemplar-based features for speech recognition towards thousands of hours of unsupervised, noisy data. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Kyoto, Japan, 2012), pp. 4437–4440. pages 90, 114

[71] Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am. 57*, 4 (Apr. 1990), 1738–52. pages 15

[72] Hermansky, H., and Morgan, N. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing 2*, 4 (Oct 1994), 578–589. pages 15

[73] Hermus, K., Wambacq, P., and Van hamme, H. A review of signal subspace speech enhancement and its application to noise robust speech recognition. *EURASIP Journal on Advances in Signal Processing*, 1 (2007), 1–15. pages 16

[74] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine 29*, 6 (Nov 2012), 82–97. pages 7

[75] Hinton, G. E., Osindero, S., and Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput. 18*, 7 (July 2006), 1527–1554. pages 7

[76] Hirsch, H., and Ehrlicher, C. Noise estimation techniques for robust speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1995), pp. 153–156. pages 136

[77] Hirsch, H. G., and Pearce, D. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ISCA Tutorial and Research Workshop ASR2000* (Sept. 2000), pp. 181–188. pages 29, 39, 49, 51, 66, 68, 82, 99, 118, 129, 138

[78] Hirsch, H. G., and Pearce, D. Applying the Advanced ETSI frontend to the Aurora-2 task. Tech. rep., Sept. 2006. version 1.1. pages 71, 105, 142

[79] Hoyer, P. Non-negative sparse coding. In *IEEE Workshop on Neural Networks for Signal Processing* (2002), pp. 557–565. pages 59, 81, 94, 116, 128, 138

[80] Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res. 5* (Dec. 2004), 1457–1469. pages 58, 94, 128, 137

[81] Huang, J., and Zhao, Y. Energy-constrained signal subspace method for speech enhancement and recognition. *IEEE Signal Processing Letters 4*, 10 (Oct. 1997), 283–285. pages 16

[82] Huang, X., Acero, A., and Hon, H.-W. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001. pages 3, 12, 13

[83] Huo, Q., Jiang, H., and Lee, C.-H. A bayesian predictive classification approach to robust speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Apr 1997), vol. 2, pp. 1547–1550. pages 16

[84] Hurmalainen, A., Gemmeke, J., and Virtanen, T. Non-negative matrix deconvolution in noise robust speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2011), pp. 4588–4591. pages 24, 34, 44, 56, 114, 126

[85] Hurmalainen, A., Gemmeke, J. F., and T., V. Modelling non-stationary noise with spectral factorisation in automatic speech recognition. *Computer Speech & Language 27*, 3 (2012), 763–779. pages 115

[86] Hurmalainen, A., Gemmeke, J. F., and Virtanen, T. Modelling non-stationary noise with spectral factorisation in automatic speech recognition. *Computer Speech & Language 27*, 3 (2013), 763–779. pages 63

[87] Ishizuka, K., and Nakatani, T. A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition. *Speech Communication 48*, 11 (2006), 1447 – 1457. pages 15

[88] Itakura, F. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 23*, 1 (Feb. 1975), 67–72. pages 41

[89] Jensen, J., Benesty, J., Christensen, M., and Jensen, S. Enhancement of single-channel periodic signals in the time-domain. *IEEE Transactions on Audio, Speech, and Language Processing 20*, 7 (2012), 1948–1963. pages 126

[90] Kalinli, O., Seltzer, M., Droppo, J., and Acero, A. Noise adaptive training for robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 18*, 8 (Nov 2010), 1889–1901. pages 142

[91] Kamm, T. M., and Meyer, G. G. L. Selective sampling of training data for speech recognition. In *Proc. HLT '02* (2002), pp. 20–24. pages 114

[92] Kanevsky, D., Sainath, T., Ramabhadran, B., and Nahamoo, D. An analysis of sparseness and regularization in exemplar-based methods for speech classification. In *Proc. INTERSPEECH* (Makuhari, Chiba, Japan, 2010), pp. 2842–2845. pages 80, 90, 126

[93] Katz, S. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing 35*, 3 (Mar 1987), 400–401. pages 13

[94] Kaufman, L., and Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*, 9th ed. Mar. 1990. pages 118

[95] Kim, D., Sra, S., and Dhillon, I. S. Fast Newton-type methods for the least squares nonnegative matrix approximation problem. In *Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA* (2007), SIAM, pp. 343–354. pages 149

[96] Kim, D.-S., Lee, S.-Y., and Kil, R. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE Transactions on Speech and Audio Processing 7*, 1 (Jan 1999), 55–69. pages 15

[97] King, B., Fevotte, C., and Smaragdis, P. Optimal cost function and magnitude power for NMF-based speech separation and music interpolation. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)* (2012), pp. 1–6. pages 91

[98] Kingsbury, B. E., Morgan, N., and Greenberg, S. Robust speech recognition using the modulation spectrogram. *Speech Communication 25*, 1–3 (1998), 117 – 132. pages 15

[99] Kneser, R., and Ney, H. Improved Backing-off for M-gram Language Modeling. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. I, pp. 181–184. pages 13

[100] Kolossa, D., and Haeb-Umbach, R., Eds. *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*. Springer, 2011. pages 16, 136

[101] Laurila, K. Noise robust speech recognition with state duration constraints. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Apr. 1997), vol. 2, pp. 871–874. pages 26

[102] Lee, A., Kawahara, T., and Doshita, S. An efficient two-pass search algorithm using word trellis index. In *Proc. ICSLP* (1998). pages 13

[103] Leggetter, C., and Woodland, P. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language 9*, 2 (1995), 171 – 185. pages 16

[104] Leutnant, V., Krueger, A., and Haeb-Umbach, R. Bayesian feature enhancement for reverberation and noise robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 21*, 8 (Aug 2013), 1640–1652. pages 16

[105] Li, B., and Sim, K. C. Improving robustness of deep neural networks via spectral masking for automatic speech recognition. In *Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)* (2013), pp. 279–284. pages 136, 142

[106] Li, B., and Sim, K. C. A spectral masking approach to noise-robust speech recognition using deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 22*, 8 (Aug 2014), 1296–1305. pages 142

[107] Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 22*, 4 (2014), 745–777. pages 15, 136

[108] Liao, H., and Gales, M. J. F. Joint uncertainty decoding for noise robust speech recognition. In *Proc. INTERSPEECH* (2005), pp. 3129–3132. pages 16

[109] Lippmann, R., Martin, E., and Paul, D. Multi-style training for robust isolated-word speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Apr 1987), vol. 12, pp. 705–708. pages 15

[110] Liu, F.-H., Stern, R. M., Huang, X., and Acero, A. Efficient cepstral normalization for robust speech recognition. In *Proceedings of the Workshop on Human Language Technology* (1993), HLT '93, pp. 69–74. pages 16

[111] Lockwood, P., and Boudy, J. Experiments with a nonlinear spectral subtractor (NSS), hidden markov models and the projection, for robust speech recognition in cars. *Speech Communication 11*, 2-3 (June 1992), 215–228. pages 136

[112] Loizou, P. C. *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, 1 ed. June 2007. pages 126

[113] Maas, A. L., Le, Q. V., O'Neil, T., Vinyals, O., Nguyen, P., and Ng, A. Y. Recurrent neural networks for noise reduction in robust ASR. In *Proc. INTERSPEECH* (2012), pp. 22–25. pages 136

[114] Macho, D., Mauuary, L., Noé, B., Cheng, Y. M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., and Saadoun, F. Evaluation of a noise-robust DSR front-end on aurora databases. In *Proc. INTERSPEECH* (2002), pp. 17–20. pages 136

[115] Mahoney, M. W., and Drineas, P. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences 106*, 3 (2009), 697–702. pages 44, 47, 63, 114, 120

[116] Martin, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing 9*, 5 (Jul. 2001), 504–512. pages 126

[117] Ming, J., Srinivasan, R., and Crookes, D. A corpus-based approach to speech enhancement from nonstationary noise. *IEEE Transactions on Audio, Speech, and Language Processing 19*, 4 (May 2011), 822–836. pages 126

[118] Mohammadiha, N., and Doclo, S. Single-channel dynamic exemplar-based speech enhancement. In *Proc. INTERSPEECH* (Sept. 2014), pp. 2690–2694. pages 126, 136

[119] Mohammadiha, N., Smaragdis, P., and Leijon, A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing 21*, 10 (2013), 2140–2151. pages 126

[120] Mohri, M., Pereira, F., and Riley, M. Weighted finite-state transducers in speech recognition. *Computer Speech & Language 16*, 1 (2002), 69 – 88. pages 13

[121] Moreno, P., Raj, B., Gouvea, E., and Stern, R. Multivariate-gaussian-based cepstral normalization for robust speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 1995), vol. 1, pp. 137–140. pages 16

[122] Moreno, P. J., Raj, B., and Stern, R. M. A vector Taylor series approach for environment-independent speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 1996), vol. 2, pp. 733–736. pages 16

[123] Mowlaee, P., Morales-Cordovilla, J. A., Pernkopf, F., Pessentheiner, H., Hagmuller, M., and Kubin, G. The 2nd CHIME speech separation and recognition challenge: Approaches on single-channel source separation and model-driven speech enhancement. In *The second CHiME Speech Separation and Recognition Challenge* (2013), pp. 59–64. pages 144

[124] Murveit, H., Butzberger, J., Digalakis, V., and Weintraub, M. Large-vocabulary dictation using sri's decipher speech recognition system: progressive search techniques. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (April 1993), vol. 2, pp. 319–322 vol.2. pages 13

[125] Nagroski, A., Boves, L., and Steeneken, H. In search of optimal data selection for training of automatic speech recognition systems. In *Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)* (Nov. 2003), pp. 67–72. pages 114

[126] Narayanan, A., and Wang, D. Investigation of speech separation as a front-end for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 22*, 4 (Apr. 2014), 826–835. pages 136

[127] Nesta, F., Matassoni, M., and Astudillo, R. F. A flexible spatial blind source extraction framework for robust speech recognition in noisy environments. In *The second CHiME Speech Separation and Recognition Challenge* (2013), pp. 33–38. pages 144

[128] Ney, H. The use of a one-stage dynamic programming algorithm for connected word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 32*, 2 (Apr 1984), 263–271. pages 14, 60, 95, 96, 117

[129] Ney, H., and Ortmanns, S. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine 16*, 5 (Sep 1999), 64–83. pages 90

[130] Olszewski, D., and Šter, B. Asymmetric clustering using the alpha-beta divergence. *Pattern Recognition 47*, 5 (2014), 2031–2041. pages 150

[131] Ortmanns, S., Ney, H., and Aubert, X. A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech & Language 11*, 1 (1997), 43 – 72. pages 13

[132] Pikrakis, A., Theodoridis, S., and Kamarotos, D. Recognition of isolated musical patterns using context dependent dynamic time warping. *IEEE Transactions on Speech and Audio Processing 11*, 3 (May 2003), 175–183. pages 34

[133] Rabiner, L. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2 (Feb. 1989), 257–286. pages 7

[134] Rabiner, L., Rosenberg, A., and Levinson, S. Considerations in dynamic time warping algorithms for discrete word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 26*, 6 (Dec 1978), 575–582. pages 114

[135] Rabiner, L., Wilpon, J., and Juang, B. A model-based connected-digit recognition system using either hidden Markov models or templates. *Computer Speech & Language 1*, 2 (1986), 167–197. pages 55

[136] Raj, B., Seltzer, M. L., and Stern, R. M. Reconstruction of missing features for robust speech recognition. *Speech Communication 43*, 4 (2004), 275 – 296. pages 16

[137] Raj, B., and Stern, R. Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine 22*, 5 (Sept. 2005), 101–116. pages 16

[138] Raj, B., Virtanen, T., Chaudhuri, S., and Singh, R. Non-negative matrix factorization based compensation of music for automatic speech recognition. In *Proc. INTERSPEECH* (Makuhari, Chiba, Japan, 2010), pp. 717–720. pages 59, 80, 91

[139] Rix, A., Beerends, J., Hollier, M., and Hekstra, A. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2001), pp. 749–752. pages 130

[140] Sainath, T. N., Chung, I., Ramabhadran, B., Picheny, M., Gunnels, J. A., Kingsbury, B., Saon, G., Austel, V., and

Chaudhari, U. V. Parallel deep neural network training for LVCSR tasks using blue gene/q. In *Proc. INTERSPEECH* (2014), pp. 1048–1052. pages 9

[141] Sainath, T. N., Nahamoo, D., Kanevsky, D., Ramabhadran, B., and Shah, P. A convex hull approach to sparse representations for exemplar-based speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (Hawaii, USA, Dec. 2011), pp. 59–64. pages 52

[142] Sainath, T. N., Ramabhadran, B., Nahamoo, D., Kanevsky, D., and Sethy, A. Sparse representations features for speech recognition. In *Proc. INTERSPEECH* (Sept. 2010), pp. 2254–2257. pages 24, 34, 44, 56, 90, 114

[143] Sainath, T. N., Ramabhadran, B., Nahamoo, D., Kanevsky, D., Van Compernolle, D., Demuynck, K., Gemmeke, J. F., Bellegarda, J. R., and Sundaram, S. Exemplar-based processing for speech recognition: An overview. *IEEE Signal Processing Magazine 29*, 6 (Nov. 2012), 98–113. pages 34, 44, 55, 80, 90, 114

[144] Sakoe, H., and Chiba, S. A dynamic programming approach to continuous speech recognition. In *Proceedings of the Seventh International Congress on Acoustics* (Budapest, Hungary, 1971), vol. 3, pp. 65–69. pages 55, 90, 114

[145] Sakoe, H., and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing 26*, 1 (Feb. 1978), 43–49. pages 10, 34, 38

[146] Saon, G., Huerta, J. M., and Jan, E.-E. Robust digit recognition in noisy environments: the ibm aurora 2 system. In *Proc. INTERSPEECH* (2001), P. Dalsgaard, B. Lindberg, H. Benner, and Z.-H. Tan, Eds., pp. 629–632. pages 16

[147] Seltzer, M. L., Dong, Y., and Wang, Y. An investigation of deep neural networks for noise robust speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (May 2013), pp. 7398–7402. pages 136

[148] Senior, A. W., Heigold, G., Bacchiani, M., and Liao, H. GMM-free DNN acoustic model training. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2014), pp. 5602–5606. pages 9

[149] Seppi, D., and Van Compernolle, D. Data pruning for template-based automatic speech recognition. In *Proc. INTERSPEECH* (Makuhari, Chiba, Japan, Sept. 2010), pp. 985–988. pages 44, 63, 90, 114

[150] Siohan, O., Chesta, C., and Lee, C.-H. Joint maximum a posteriori adaptation of transformation and HMM parameters. *IEEE Transactions on Speech and Audio Processing 9*, 4 (May 2001), 417–428. pages 16

[151] Smaragdis, P. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 1 (2007), 1–12. pages 59, 91, 126

[152] Smaragdis, P., and Brown, J. C. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2003), pp. 177–180. pages 59, 91

[153] Smaragdis, P., Shashanka, M., and Raj, B. A sparse non-parametric approach for single channel separation of known sounds. In *NIPS* (2009), pp. 1705–1713. pages 126, 136

[154] Sreenivas, T., and Kirnapure, P. Codebook constrained Wiener filtering for speech enhancement. *IEEE Transactions on Speech and Audio Processing 4*, 5 (1996), 383–389. pages 126

[155] Srinivasan, S., and Wang, D. Transforming binary uncertainties for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 7 (Sept 2007), 2130–2140. pages 16

[156] Stouten, V., Van hamme, H., and Wambacq, P. Model-based feature enhancement with uncertainty decoding for noise robust {ASR}. *Speech Communication 48*, 11 (2006), 1502 – 1514. pages 16

[157] Sun, X., Chen, X., and Zhao, Y. On the effectiveness of statistical modeling based template matching approach for continuous speech recognition. In *Proc. INTERSPEECH* (2011), pp. 453–456. pages 55

[158] Sun, X., and Zhao, Y. New methods for template selection and compression in continuous speech recognition. In *Proc. INTERSPEECH* (Florence, Italy, Aug. 2011), pp. 985–988. pages 44, 63, 90, 114

[159] Sun, Y., Gemmeke, J. F., Cranen, B., ten Bosch, L., and Boves, L. Fusion of parametric and non-parametric approaches to noise-robust {ASR}. *Speech Communication 56*, 0 (2014), 49 – 62. pages 90, 114

[160] Sundaram, S., and Bellegarda, J. R. Latent perceptual mapping: a new acoustic modeling framework for speech recognition. In *Proc. INTERSPEECH* (Sept. 2010), pp. 881–884. pages 55

[161] Sundaram, S., and Bellegarda, J. R. Latent perceptual mapping with data-driven variable-length acoustic units for template-based speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2012), pp. 4125–4128. pages 90, 114

[162] Tan, Q. F., and Narayanan, S. S. Novel variations of group sparse regularization techniques with applications to noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing 20*, 4 (May 2012), 1337–1346. pages 24, 34, 44, 56, 59, 90, 91, 114, 126

[163] Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society (Series B) 58* (1996), 267–288. pages 59

[164] Tran, D. T., Vincent, E., Jouvet, D., and Adiloglu, K. Using full-rank spatial covariance models for noise-robust ASR. In *The second CHiME Speech Separation and Recognition Challenge* (2013), pp. 31–32. pages 144, 145

[165] Trentin, E., and Gori, M. A survey of hybrid ANN/HMM models for automatic speech recognition. *Neurocomputing 37*, 1–4 (2001), 91 – 126. pages 7

[166] Tsao, Y., Li, J., Lee, C.-H., and Nakamura, S. Soft margin estimation on improving environment structures for ensemble speaker and speaking environment modeling. In *Proc. 3rd Int. Universal Communication Sym.*, pp. 404–408. pages 142

[167] Van Compernolle, D. Noise adaptation in a hidden markov model speech recognition system. *Computer Speech & Language 3*, 2 (1989), 151–167. pages 136

[168] Van hamme, H. PROSPECT features and their application to missing data techniques for robust speech recognition. In *Proc. INTERSPEECH* (2004), pp. 101–104. pages 15

[169] Van hamme, H. A diagonalized Newton algorithm for non-negative sparse coding. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2013), IEEE, pp. 7299–7303. pages 149

[170] Van Segbroeck, M., and Van hamme, H. Advances in missing feature techniques for robust large-vocabulary continuous speech recognition.

IEEE Transactions on Audio, Speech, and Language Processing 19, 1 (Jan. 2011), 123–137. pages 16

[171] Varga, A., and Moore, R. Hidden markov model decomposition of speech and noise. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Apr 1990), pp. 845–848 vol.2. pages 16

[172] Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., and Matassoni, M. The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Vancouver, Canada, May 2013), pp. 126–130. pages 65, 83, 99, 119, 139, 140, 144

[173] Vincent, E., Gribonval, R., and Fevotte, C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing 14*, 4 (2006), 1462–1469. pages 130

[174] Virtanen, T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing 15*, 3 (March 2007), 1066–1074. pages 27, 58, 59, 81, 91, 94, 126, 128, 137, 138

[175] Virtanen, T., Gemmeke, J. F., and Raj, B. Active-set Newton algorithm for overcomplete non-negative representations of audio. *IEEE Transactions on Audio, Speech & Language Processing 21*, 11 (2013), 2277–2289. pages 149

[176] Virtanen, T., Singh, R., and Raj, B. *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012. pages 55, 136

[177] Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13*, 2 (April 1967), 260–269. pages 13

[178] White, G., and Neely, R. Speech recognition experiments with linear predication, bandpass filtering, and dynamic programming. *IEEE Transactions on Acoustics, Speech and Signal Processing 24*, 2 (Apr 1976), 183–188. pages 114

[179] Wu, Y., Zhang, R., and Rudnicky, A. Data selection for speech recognition. In *Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)* (Dec. 2007), pp. 562–565. pages 114

[180] Yapanel, U. H., and Hansen, J. H. A new perceptually motivated mvdr-based acoustic front-end (pmvdr) for robust automatic speech recognition. *Speech Communication 50*, 2 (2008), 142 – 152. pages 15

[181] Yilmaz, E., Gemmeke, J. F., Van Compernolle, D., and Van hamme, H. Noise-robust digit recognition with exemplar-based sparse representations of variable length. In *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–4. pages 34, 44, 51, 56, 64, 68, 80

[182] Yilmaz, E., Gemmeke, J. F., and Van hamme, H. Exemplar selection techniques for sparse representations of speech using multiple dictionaries. In *European Signal Processing Conference (EUSIPCO)* (Marrakesh, Morocco), pp. 1–5. pages 56, 63, 64, 82, 98, 115, 120, 122, 123

[183] Yilmaz, E., Gemmeke, J. F., and Van hamme, H. Noise-robust automatic speech recognition with exemplar-based sparse representations using multiple length adaptive dictionaries. In *2nd International Workshop on Machine Learning in Multisource Environments (CHIME)* (Vancouver, Canada, June 2013), pp. 39–43. pages 56, 63, 67, 97, 120, 140

[184] Yilmaz, E., Gemmeke, J. F., and Van hamme, H. Noise robust exemplar matching using sparse representations of speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(8)* (Aug. 2014), 1306–1319. pages 80, 81, 83, 84, 91, 92, 96, 97, 98, 100, 102, 104, 107, 115, 119, 126, 128, 130, 136, 138, 139, 140, 141, 142, 143

[185] Yilmaz, E., Gemmeke, J. F., and Van hamme, H. Noise-robust speech recognition with exemplar-based sparse representations using alpha-beta divergence. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Florence, Italy, May 2014), pp. 5539–5543. pages 91, 102, 111

[186] Yilmaz, E., Gemmeke, J. F., and Van hamme, H. Noise robust exemplar matching with alpha-beta divergence. *Submitted to Speech Communication* (2015). pages 115, 117, 119, 120, 136

[187] Yilmaz, E., Van Compernolle, D., and Van hamme, H. Combining exemplar-based matching and exemplar-based sparse representations of speech. In *Symposium on Machine Learning in Speech and Language Processing (MLSLP)* (Portland, OR, USA, Sept. 2012). pages 34, 35, 38, 39, 40, 44, 45, 49, 56, 61

[188] Yu, D., and Deng, L. *Automatic Speech Recognition - A Deep Learning Approach*. Springer, 2014. pages 7, 9

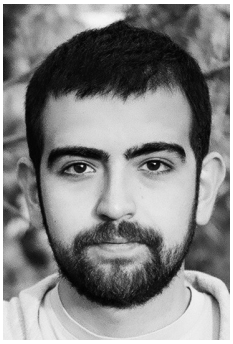
[189] Yu, D., Deng, L., Droppo, J., Wu, J., Gong, Y., and Acero, A. Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor. *IEEE Transactions on Audio, Speech, and Language Processing 16*, 5 (July 2008), 1061–1070. pages 136

[190] Zhang, C., and Woodland, P. C. Standalone training of context-dependent deep neural network acoustic models. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2014), pp. 5597–5601. pages 9

[191] Zhou, P., Dai, L., and Jiang, H. Sequence training of multiple deep neural networks for better performance and faster training speed. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2014), pp. 5627–5631. pages 9

[192] Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society 67* (2005), 301–320. pages 59

Short Biography



Emre Yılmaz was born on 15 September 1986 in Konya, Turkey. He received the B.Sc. degree in electrical and electronics engineering from the Middle East Technical University (METU), Turkey in 2008 and the M.Sc. degree in electrical engineering from the Royal Institute of Technology (KTH), Sweden in 2010. Then, he worked as a part-time researcher in the Institute of Communication Systems and Data Processing (IND), RWTH Aachen, Germany. He joined the Department of Electrical Engineering (ESAT), KU Leuven, Belgium as a Ph.D. candidate in January, 2011 and received the Ph.D. degree in May, 2015. His research interests are noise robust automatic speech recognition, recognition of children speech and medical applications of automatic speech recognition.

List of Publications

Articles in International Journals

1. **Emre Yılmaz**, Jort F. Gemmeke and Hugo Van hamme, "*Noise Robust Exemplar Matching Using Sparse Representations of Speech*", IEEE/ACM Transactions on Audio, Speech, and Language Processing, volume 22, No. 8, pages 1306-1319, Aug. 2014.

2. **Emre Yılmaz**, Jort F. Gemmeke and Hugo Van hamme, "*Noise Robust Exemplar Matching with Alpha-Beta Divergence*", Submitted to Speech Communication, 2015.

Articles in International Conferences

1. Heiner Löllmann, **Emre Yılmaz**, Marco Jeub, and Peter Vary, "*An Improved Algorithm for Blind Reverberation Time Estimation*", In Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC), pages 1-4, Tel Aviv, Israel, August 2010.

2. **Emre Yılmaz**, Jort F. Gemmeke, Dirk Van Compernolle and Hugo Van hamme, "*Noise-robust Digit Recognition with Exemplar-based Sparse Representations of Variable Length*", In IEEE Workshop on Machine Learning for Signal Processing (MLSP), pages 23-26, Santander, Spain, September 2012.

3. **Emre Yılmaz**, Dirk Van Compernolle and Hugo Van hamme, "*Combining Exemplar-based Matching and Exemplar-based Sparse Representations of Speech*", In Symposium on Machine Learning in Speech and Language Processing (MLSLP), Portland, USA, September 2012.

4. **Emre Yılmaz**, Dirk Van Compernolle and Hugo Van hamme, "*Robust Tracking for Automatic Reading Tutors*", In Proc. INTERSPEECH, pages 1-4, Portland, USA, September 2012.

5. **Emre Yılmaz**, Jort F. Gemmeke and Hugo Van hamme, "*Embedding Time Warping in Exemplar-based Sparse Representations of Speech*", In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 8076-8080, Vancouver, Canada, May 2013.

6. **Emre Yılmaz**, Jort F. Gemmeke and Hugo Van hamme, "*Noise-robust Automatic Speech Recognition with Exemplar-based Sparse Representations Using Multiple Length Adaptive Dictionaries*", In 2nd International Workshop on Machine Learning in Multisource Environments (CHIME), pages 39-43, Vancouver, Canada, June 2013.

7. Hanne Deprez, **Emre Yılmaz**, Stefan Lievens and Hugo Van hamme, "*Automating Speech Reception Threshold Measurements Using Automatic Speech Recognition*", In 4th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), pages 1-6, Grenoble, France, August 2013.

8. **Emre Yılmaz**, Jort F. Gemmeke and Hugo Van hamme, "*Exemplar Selection Techniques for Sparse Representations of Speech Using Multiple Dictionaries*", In 21st European Signal Processing Conference (EUSIPCO), pages 1-5, Marrakesh, Morocco, Sept. 2013.

9. **Emre Yılmaz**, Jort F. Gemmeke and Hugo Van hamme, "*Noise-robust Speech Recognition with Exemplar-based Sparse Representations Using Alpha-Beta Divergence*", In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 5539-5543, Florence, Italy, May 2014.

10. **Emre Yılmaz**, Joris Pelemans, Stefan Lievens and Hugo Van hamme, "*Speech Reception Threshold Measurement Using Automatic Speech Recognition*", In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 1-4, Show & Tell Session, Florence, Italy, May 2014.

11. **Emre Yılmaz**, Joris Pelemans and Hugo Van hamme, "*Automatic Assessment of Children's Reading with the FLaVoR Decoding Using a Phone Confusion Model*", In Proc. INTERSPEECH, pages 969-972, Singapore, Sept. 2014.

12. **Emre Yılmaz**, Konstantinos Rematas, Tinne Tuytelaars and Hugo Van hamme, "*Learning Like a Toddler: Watching Television Series to Learn*

Vocabulary from Images and Audio", In the 22nd ACM International Conference on Multimedia, pages 1189-1192, Orlando, Florida, USA, Nov. 2014.

13. **Emre Yılmaz**, Deepak Baby and Hugo Van hamme, "*Noise Robust Exemplar Matching with Coupled Dictionaries for Single-Channel Speech Enhancement*", Submitted to EUSIPCO 2015.

14. **Emre Yılmaz**, Jort F. Gemmeke and Hugo Van hamme, "*Adaptive Noise Dictionary Design for Noise Robust Exemplar Matching of Speech*", Submitted to EUSIPCO 2015.

15. **Emre Yılmaz**, Jort F. Gemmeke and Hugo Van hamme, "*Data Selection for Noise Robust Exemplar Matching*", Submitted to INTERSPEECH 2015.

16. **Emre Yılmaz**, Deepak Baby and Hugo Van hamme, "*Noise Robust Exemplar Matching for Speech Enhancement: Applications to Automatic Speech Recognition*", Submitted to INTERSPEECH 2015.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING (ESAT)
CENTER FOR PROCESSING SPEECH AND IMAGES (PSI)
Kasteelpark Arenberg 10
B-3001 Heverlee
emre.yilmaz@esat.kuleuven.be
http://www.esat.kuleuven.be